SIAM J. NUMER. ANAL.
Vol. 58, No. 2, pp. 1263–1294

TIKHONOV REGULARIZATION WITHIN ENSEMBLE KALMAN INVERSION*

NEIL K. CHADA†, ANDREW M. STUART‡, AND XIN T. TONG§

Abstract. Ensemble Kalman inversion is a parallelizable methodology for solving inverse or parameter estimation problems. Although it is based on ideas from Kalman filtering, it may be viewed as a derivative-free optimization method. In its most basic form it regularizes ill-posed inverse problems through the subspace property: the solution found is in the linear span of the initial ensemble employed. In this work we demonstrate how further regularization can be imposed, incorporating prior information about the underlying unknown. In particular we study how to impose Tikhonov-like Sobolev penalties. As well as introducing this modified ensemble Kalman inversion methodology, we also study its continuous-time limit, proving ensemble collapse; in the language of multi-agent optimization this may be viewed as reaching consensus. We also conduct a suite of numerical experiments to highlight the benefits of Tikhonov regularization in the ensemble inversion context.

Key words. Tikhonov regularization, ensemble Kalman inversion, Bayesian inverse problems, long-term behavior

AMS subject classifications. 35Q93, 58E25, 65F22, 65M32

DOI. 10.1137/19M1242331

1. Introduction.

Inverse problems are ubiquitous in science and engineering. They occur in numerous applications, such as recovering permeability from measurement of flow in a porous medium [31, 35] or locating pathologies via medial imaging [23]. Mathematically speaking, an inverse problem may be formulated as the recovery of parameter $u \in X$ from noisy data $y \in Y$ where the parameter u and data y are related by

$$y = G(u) + \eta, \tag{1.1}$$

G is an operator from the space of parameters to observations, and η represents noise; in this paper we will restrict ourselves to X, Y being separable Hilbert spaces. Inverse problems are typically solved through two competing methodologies: the deterministic optimization approach [13] and the probabilistic Bayesian approach [23]. The former is based on defining a loss function $\ell(G(u), y)$ which one aims to minimize; a regularizer $R(u)$ that incorporates prior information about u is commonly added to improve the inversion [3]. The Bayesian approach instead views u, y, and η as random variables and focuses on the conditional distribution of $u|y$ via Bayes's theorem as the

*Received by the editors February 6, 2019; accepted for publication (in revised form) January 30, 2020; published electronically April 22, 2020.
https://doi.org/10.1137/19M1242331
Funding: The first author's research was supported by Singapore Ministry of Education Academic Research Funds Tier 2 grant MOE2016-T2-2-135. The second author's research was supported by Office of Naval Research grant 00014-17-1-2079 and National Science Foundation grant DMS-1818977. The third author's research was supported by Singapore Ministry of Education Tier 1 grant R-146-000-292-114.
†Department of Applied Probability and Statistics, National University of Singapore, Singapore, 117546, Singapore (neil.chada@nus.edu.sg).
‡Department of Computing and Mathematical Sciences, Caltech, Pasadena, CA 91125 (astuart@caltech.edu).
§Department of Mathematics, National University of Singapore, Singapore, 117543, Singapore (mattxin@nus.edu.sg).

solution; this approach has received recent attention since it provides representation of the underlying uncertainty, and it may be formulated even in the infinite-dimensional setting [41]. The optimization and Bayesian approaches are linked via the notion of the maximum a posteriori (MAP) estimator through which the mode of the conditional distribution on $u|y$ is shown to correspond to optimization of a regularized loss function [1, 9, 18, 23, 32].

Ensemble Kalman inversion (EKI) is a proposed inversion methodology that lies at the interface between the deterministic and probabilistic approaches [5, 21]. It is based on the ensemble Kalman filter (EnKF) [15, 16, 28, 37], which is an algorithm originally designed for high-dimensional state estimation, derived by combining sequential Bayesian methods with an approximate Gaussian ansatz. EKI applies EnKF to the inverse problem setting by introducing a trivial dynamics for the unknown. The idea of using ensemble Kalman methods for inverse problems was pioneered in the oil-reservoir simulation community [35] and, in particular, the idea of iterating using trivial dynamics conditioned on data was introduced in [8, 12]. The algorithm works by iteratively updating an ensemble of candidate solutions $\{u_n^{(j)}\}_{j=1}^J$ from iteration index n to $n+1$; here j indexes the ensemble and J denotes the size of the ensemble. The basic form of the algorithm is as follows. Define the empirical means

$$\bar{u}_n = \frac{1}{J}\sum_{j=1}^{J} u_n^{(j)}, \quad \bar{G}_n = \frac{1}{J}\sum_{j=1}^{J} G(u_n^{(j)})$$

and covariances

$$C_n^{uu} = \frac{1}{J}\sum_{j=1}^{J}\big(u_n^{(j)} - \bar{u}_n\big) \otimes \big(u_n^{(j)} - \bar{u}_n\big), \tag{1.2a}$$

$$C_n^{up} = \frac{1}{J}\sum_{j=1}^{J}\big(u_n^{(j)} - \bar{u}_n\big) \otimes \big(G\big(u_n^{(j)}\big) - \bar{G}_n\big), \tag{1.2b}$$

$$C_n^{pp} = \frac{1}{J}\sum_{j=1}^{J}\big(G(u_n^{(j)}) - \bar{G}_n\big) \otimes \big(G(u_n^{(j)}) - \bar{G}_n\big). \tag{1.2c}$$

Then the EKI update formulae are

$$u_{n+1}^{(j)} = u_n^{(j)} + C_n^{up}\big(C_n^{pp} + \Gamma\big)^{-1}\big(y_{n+1}^{(j)} - G(u_n^{(j)})\big), \tag{1.3}$$

where the artificial observations are given by

$$y_{n+1}^{(j)} = y + \xi_{n+1}^{(j)}, \quad \xi_{n+1}^{(j)} \sim \mathcal{N}(0, \Gamma') \quad \text{i.i.d.} \tag{1.4}$$

Here an implicit assumption is that η is additive centered Gaussian noise with covariance Γ and it is independent of u. Typical choices for Γ' include $\mathbf{0}$ and Γ. The history of the development of the method, which occurred primarily within the oil industry, may be found in [35]; the general and application-neutral formulation of the method as presented here may be found in [21].

For linear, bounded, and invertible G the method provably optimizes the standard least squares loss function over the finite-dimensional subspace spanned by the initial ensemble differences [39, Theorem 4.3, Corollary 4.4]; for nonlinear G similar behavior is observed empirically in [21]. However, the ensemble does not, in general, accurately

capture posterior variability; this is demonstrated theoretically in [14, Lemma 12, Theorem 13] and numerically in [21, 27]. For this reason we focus on the perspective of EKI as a derivative-free optimization method, somewhat similar in spirit to the paper [45] concerning the EnKF for state estimation. Viewed in this way, EKI may be seen as part of a wider class of tools based around multi-agent interacting systems which aim to optimize via consensus [36]. Within this context of EKI as an optimization tool for inversion, a potential drawback is the issue of how to incorporate regularization. It is demonstrated in [21, 31] that the updated ensemble lies within the linear span of the initial ensemble and this is a form of regularization since it restricts the solution to a finite-dimensional space. However, the numerical evidence in [21] demonstrates that overfitting may still occur, and this led to the imposition of iterative regularization by analogy with the Levenburg–Marquardt approach, a method pioneered in [19]; see [5] for an application of this approach.

There are a number of approaches to regularization of ill-posed inverse problems which are applied in the deterministic optimization realm. Three primary ones are (i) optimization over a compact set, (ii) iterative regularization through early stopping, and (iii) Tikhonov penalization of the misfit. The standard EKI imposes approach (i), and the method of [19] imposes approach (ii). The purpose of this paper is to demonstrate how approach (iii), Tikhonov regularization [3, 13], may also be incorporated into the EKI approach. Our primary contributions are the following:

- We present a straightforward modification of the standard EKI methodology from [21] which allows for incorporation of Tikhonov regularization, leading to the TEKI (Tikhonov-EKI) approach.
- We study the TEKI approach analytically, building on the continuous time analysis and gradient flow structure for EKI developed in [39, section 3]; in particular we prove that, for general nonlinear inverse problems, the TEKI flow exhibits asymptotic consensus, i.e., ensemble collapse as the iteration count tends to infinity.
- We describe numerical experiments which highlight the benefits of TEKI over EKI, using inverse problems arising from the eikonal equation [11]. We further test our methodology on Darcy flow [21] to highlight the robustness of the proposed algorithm.

The outline of the paper is as follows. In section 2 we describe the TEKI methodology, introducing the modified inverse problem which incorporates the additional regularization. Section 3 is devoted to the derivation of a continuous time analogue of the resulting algorithm, and we also study its properties in the case of linear inverse problems. In section 4 we present numerical experiments demonstrating the benefits of using TEKI over EKI, using inverse problems arising from the eikonal equation and Darcy flow. We conclude in section 5 with an overview and further research directions to consider.

2. EKI with Tikhonov regularization. In this section we derive the TEKI algorithm, the regularized variant of the EKI algorithm which we introduce in this paper. We start by recalling how classical Tikhonov regularization works and then demonstrate how to apply similar ideas within EKI.

Assuming that we model $\eta \sim N(0, \Gamma)$ in (1.1), the resulting loss function is in the L^2 form

$$\ell_Y(y', y) = \frac{1}{2}\|\Gamma^{-1/2}(y' - y)\|_Y^2. \tag{2.1}$$

Recall (see the previous section) that EKI minimizes

$$\ell_Y(G(u), y) = \frac{1}{2}\|\Gamma^{-1/2}(G(u) - y)\|_Y^2 \tag{2.2}$$

within a subspace defined by the initial ensemble, provably in the linear case and with similar behavior observed empirically in the nonlinear case.

Tikhonov regularization is associated with defining

$$R(u) = \frac{\boldsymbol{\lambda}}{2}\|u\|_K^2, \tag{2.3}$$

where K is a Hilbert space which is continuously and compactly embedded into X, and which minimizes the sum of $\ell(G(u), y)$ and $R(u)$. The regularization parameter $\boldsymbol{\lambda} > 0$ may be tuned to trade-off between data fidelity and smoothness, thereby avoiding overfitting. This may be connected to Bayesian regularization if the prior on u is the Gaussian measure $N(0, \boldsymbol{\lambda}^{-1}C_0)$, with C_0 trace-class and strictly positive definite on X. Then K is a Hilbert space K equipped with inner product $\langle C_0^{-\frac{1}{2}}\cdot, C_0^{-\frac{1}{2}}\cdot\rangle_X$ and norm $\|\cdot\|_K = \|C_0^{-\frac{1}{2}}\cdot\|_X$; it is known as the Cameron–Martin space associated with the Gaussian prior. Minimizing the sum of $\ell(G(u), y)$ and $R(u)$ corresponds to finding a mode of the distribution [9].

To incorporate such prior information into the EKI algorithm, we proceed as follows. We first extend (1.1) to the equations

$$y = G(u) + \eta_1, \tag{2.4a}$$

$$0 = u + \eta_2, \tag{2.4b}$$

where η_1, η_2 are independent random variables distributed as $\eta_1 \sim N(0, \Gamma)$, $\eta_2 \sim N(0, \boldsymbol{\lambda}^{-1}C_0)$. Let $Z = Y \times X$; we then define the new variables z, η and mapping $F : X \times X \to Z$ as follows:

$$z = \begin{bmatrix} y \\ 0 \end{bmatrix}, \quad F(u) = \begin{bmatrix} G(u) \\ u \end{bmatrix}, \quad \eta = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix},$$

noting that then

$$\eta \sim N(0, \Sigma), \quad \Sigma = \begin{bmatrix} \Gamma & 0 \\ 0 & \boldsymbol{\lambda}^{-1}C_0 \end{bmatrix}.$$

We then consider the inverse problem

$$z = F(u) + \eta, \tag{2.5}$$

which incorporates the original equation (1.1) via (2.4a) and the prior information via (2.4b). We now define the ensemble mean

$$\bar{F}_n = \frac{1}{J}\sum_{j=1}^{J} F(u_n^{(j)})$$

and covariances

$$B_n^{up} = \frac{1}{J}\sum_{j=1}^{J}\big(u_n^{(j)} - \bar{u}_n\big)\otimes\big(F\big(u_n^{(j)}\big) - \bar{F}_n\big), \quad B_n^{pp} = \frac{1}{J}\sum_{j=1}^{J}\big(F(u_n^{(j)}) - \bar{F}_n\big)\otimes\big(F(u_n^{(j)}) - \bar{F}_n\big).$$

The TEKI update formulae are then found by applying the EKI algorithm to (2.5) to obtain

$$u_{n+1}^{(j)} = u_n^{(j)} + B_n^{up}\big(B_n^{pp} + \Sigma\big)^{-1}\big(z_{n+1}^{(j)} - F(u_n^{(j)})\big), \tag{2.6}$$

where

$$z_{n+1}^{(j)} = z + \zeta_{n+1}^{(j)}, \quad \zeta_{n+1}^{(j)} \sim \mathcal{N}(0, \Sigma') \quad \text{i.i.d.} \tag{2.7}$$

Typical choices for Σ' are 0 and Σ. Notice that the resulting L^2 loss function (2.1) is, in this case,

$$\ell_Z(z', z) = \frac{1}{2}\|\Sigma^{-1/2}(z' - z)\|_Z^2, \tag{2.8}$$

leading, with $z' = F(u)$, to the loss function

$$\mathcal{I}(u; y) := \frac{1}{2}\|\Gamma^{-1/2}(y - G(u))\|_X^2 + \frac{\boldsymbol{\lambda}}{2}\|u\|_K^2. \tag{2.9}$$

It is in this sense that TEKI regularizes EKI, the latter being associated with the unregularized objective function (2.2).

Remark 2.1. Both the EKI algorithm (1.3) and the TEKI algorithm (2.6) have the property that all ensemble members remain in the linear span of the initial ensemble for all time. This is proved in [39] for EKI; the proof for TEKI is very similar and hence not given. For EKI (resp., TEKI) it follows simply from the fact that C_n^{up} (resp., B_n^{up}) projects onto the linear span of the current ensemble and then uses an induction.

3. Continuous time limit of TEKI. In this section we aim to study the use of Tikhonov regularization within EKI through analysis of a continuous time limit of TEKI. For economy of notation we assume the regularization constant $\boldsymbol{\lambda}$ to take the value 1 throughout. This incurs no loss of generality, since one can always replace $(\boldsymbol{\lambda}, C_0)$ with $(1, \boldsymbol{\lambda}^{-1}C_0)$, and the TEKI formulation remains the same.

In subsection 3.1 we derive the continuous time limit of the TEKI algorithm, while in subsection 3.2 we state and prove the general existence theorem (Theorem 3.1) for the TEKI flow. In subsection 3.3 we demonstrate ensemble collapse of the TEKI flow, Theorem 3.3; this shows that the ensemble members reach consensus. We also prove two lemmas which together characterize an invariant subspace property of TEKI flow, closely related to Remark 2.1. Subsection 3.4 contains derivation of two a priori bounds on the TEKI flow, one in the linear setting and the other in the general setting. Finally, in subsection 3.5 we study the long-time behavior of TEKI flow in the linear setting, generalizing related work on the EKI flow in [39].

3.1. Derivation of continuous time limit. We first recall the derivation of the continuous time limit of the EKI algorithm (1.3) from [39], as the derivation for TEKI is very similar. For this purpose we set $\Gamma' = 0$ and rescale $\Gamma \mapsto h^{-1}\Gamma$ so that (approximately for $h \ll 1$) $(C_n^{pp} + \Gamma)^{-1} \mapsto h\Gamma^{-1}$. We then view $u_n^{(j)}$ as an approximation of a continuous function $u^{(j)}(t)$ at time $t = nh$ and let $h \to 0$. To write down the resulting flow succinctly, we let $\mathfrak{u} \in X^J$ denote the collection of $\{u^{(j)}\}_{j\in\{1,\dots,J\}}$. Now define

$$D_{jk}(\mathfrak{u}) := \langle \Gamma^{-1/2}(G(u^{(j)}) - y), \Gamma^{-1/2}(G(u^{(k)}) - \bar{G})\rangle_Y,$$

where

$$\bar{u} := \frac{1}{J}\sum_{m=1}^{J} u^{(m)}, \quad \bar{G} := \frac{1}{J}\sum_{m=1}^{J} G(u^{(m)}).$$

The continuum limit of (1.3) is then

$$\begin{aligned}\frac{du^{(j)}}{dt} &= -\frac{1}{J}\sum_{k=1}^{J}(u^{(k)}-\bar{u})\otimes(G(u^{(k)})-\bar{G})\Gamma^{-1}(G(u^{(j)})-y)\\ &= -\frac{1}{J}\sum_{k=1}^{J} D_{jk}(\mathfrak{u})(u^{(k)}-\bar{u}) = -\frac{1}{J}\sum_{k=1}^{J} D_{jk}(\mathfrak{u})u^{(k)}.\end{aligned} \tag{3.1}$$

Here we used the fact that replacing $u^{(k)}(t)$ by $u^{(k)}(t)-\bar{u}(t)$ does not change the flow since $D_{jk}(\mathfrak{u}(t))$ sums to zero over k; we will use this fact occasionally in what follows and without further comment. The equations may be written as

$$\frac{d\mathfrak{u}}{dt} = -\frac{1}{J}D(\mathfrak{u})\mathfrak{u} \tag{3.2}$$

for the appropriate Kronecker operator $D(\mathfrak{u}) \in \mathcal{L}(X^J, X^J)$ defined from the $D_{jk}(\mathfrak{u})$. Note also that we hid the dependence on time t in our derivation above, and we will often do so in the discussion below.

The resulting flow is insightful because it demonstrates that, in the linear case $G(\cdot) = A\cdot$, each ensemble member undergoes a gradient flow for the loss function (2.1) preconditioned by the empirical covariance $C(\mathfrak{u})$ defined by

$$C(\mathfrak{u}) = \frac{1}{J}\sum_{m=1}^{J}\big(u^{(m)}-\bar{u}\big)\otimes\big(u^{(m)}-\bar{u}\big); \tag{3.3}$$

specifically, we have

$$\frac{du^{(j)}}{dt} = -C(\mathfrak{u})\nabla_u \ell_Y(Au^{(j)}, y). \tag{3.4}$$

Note that although each ensemble member performs a gradient flow, they are coupled through the empirical covariance.

We now carry out a similar derivation for the TEKI algorithm; doing so will demonstrate explicitly that the method introduces a Tikhonov regularization. Consider the TEKI algorithm (2.6), setting $\Sigma' = 0$, rescaling $\Sigma \mapsto h^{-1}\Sigma$, and viewing $u_n^{(j)}$ as an approximation of a continuous function $u^{(j)}(t)$ at time $t = nh$. The limiting

flow is

$$\begin{aligned}\frac{du^{(j)}}{dt} &= -\frac{1}{J}\sum_{k=1}^{J}\big(u^{(k)}-\bar{u}\big)\otimes\big(F\big(u^{(k)}\big)-\bar{F}\big)\Sigma^{-1}\big(F(u^{(j)})-z\big)\\ &= -\frac{1}{J}\sum_{k=1}^{J}\Big\langle\Gamma^{-1/2}(G(u^{(k)})-\bar{G}),\Gamma^{-1/2}(G(u^{(j)})-y)\Big\rangle_Y\big(u^{(k)}-\bar{u}\big)\\ &\quad -\frac{1}{J}\sum_{k=1}^{J}\langle u^{(j)},u^{(k)}-\bar{u}\rangle_K\big(u^{(k)}-\bar{u}\big)\\ &= -\frac{1}{J}\sum_{k=1}^{J}\Big(D_{jk}(\mathfrak{u})+\langle u^{(j)},u^{(k)}-\bar{u}\rangle_K\Big)(u^{(k)}-\bar{u})\\ &= -\frac{1}{J}\sum_{k=1}^{J}E_{jk}(\mathfrak{u})(u^{(k)}-\bar{u}), \qquad (3.5)\end{aligned}$$

where

$$E_{jk}(\mathfrak{u}) := D_{jk}(\mathfrak{u})+\langle u^{(j)},u^{(k)}-\bar{u}\rangle_K.$$

This may be written as

$$\frac{d\mathfrak{u}}{dt} = -\frac{1}{J}E(\mathfrak{u})\mathfrak{u} \qquad (3.6)$$

for an appropriate Kronecker matrix $E(\mathfrak{u})\in\mathcal{L}(X^J,X^J)$ defined from the $E_{jk}(\mathfrak{u})$.

We note that the flow may be written as

$$\frac{du^{(j)}}{dt} = -\frac{1}{J}\sum_{k=1}^{J}D_{jk}(\mathfrak{u})(u^{(k)}-\bar{u}) - C(\mathfrak{u})\nabla_u R(u^{(j)}), \qquad (3.7)$$

where

$$R(u) = \frac{1}{2}\|u\|_K^2. \qquad (3.8)$$

So we see explicitly that the algorithm includes a Tikhonov regularization, preconditioned by the empirical covariance $C(\mathfrak{u})$. In the linear case $G(\,\cdot\,)=A\,\cdot$, define

$$\mathcal{I}_{\text{linear}}(u;y) = \ell_Y(Au^{(j)},y)+R(u). \qquad (3.9)$$

Note that, with $\ell_Y(\cdot,\cdot)$ and $R(\cdot)$ defined by (2.1) and (3.8), this coincides with (2.9) when specialized to the linear case. In particular, we also see that the TEKI flow has the form

$$\frac{du^{(j)}}{dt} = -C(\mathfrak{u})\nabla_u\mathcal{I}_{\text{linear}}(u^{(j)};y). \qquad (3.10)$$

Each ensemble member thus undergoes a gradient flow with respect to the Tikhonov regularized least squares loss function $\mathcal{I}_{\text{linear}}(u;y)$, preconditioned by the empirical covariance of the collection of all the ensemble members.

Remark 3.1. Additive covariance inflation, as described in [39], modifies the EKI gradient flow (3.4) by addition of a fixed invertible covariance matrix to the empirical covariance. In contrast, (3.10) fixes the empirical covariance and instead modifies the objective function by addition of a regularizer.

Note that we derived the continuous time TEKI algorithm (3.6) by passing to the continuum limit; to implement an algorithm in subsection 4.1, we apply an Euler discretization (with adaptive time-step) to (3.6). Other choices of discrete time-algorithms could have been made, for example, using the original discrete time TEKI algorithm (2.6). We have chosen to use discretization of (3.6), rather than (2.6), because it results in a conceptually simpler algorithm and (relatedly) because it avoids the calculation of the empirical covariance in data space. However, working directly with (2.6) may have practical advantages in some problems and can be contemplated separately [7].

3.2. Existence for TEKI flow. Recall that the Cameron–Martin space associated with the Gaussian measure $N(0, C_0)$ on X is the domain of $C_0^{-\frac{1}{2}}$. We have the following result.

THEOREM 3.1. *Suppose the initial ensemble* $\{u^{(j)}(0)\}_{j=1}^J$ *is chosen to lie in* K *and that* $G : K \to Y$ *is* C^1. *Let* $\mathcal{A}$ *denote the linear span of* $\{u^{(j)}(0)\}_{j=1}^J$ *and* $\mathcal{A}^J$ *the* J*-fold Cartesian product of this set. Then* (3.6) *has a unique solution in* $C^1([0,T);\mathcal{A}^J)$ *for some* $T > 0$.

Remark 3.2. The same theorem may be proved for (3.2) under the milder assumptions that the initial ensemble $\{u^{(j)}(0)\}_{j=1}^J$ is chosen to lie in X itself and that $G : X \to Y$ is C^1.

Proof of Theorem 3.1. The right-hand side of (3.6) is of the form $E(\mathfrak{u})\mathfrak{u}$ and $E : \mathcal{A}^J \to \mathcal{L}(\mathcal{A}^J, \mathcal{A}^J)$. Thus it suffices to show that E is differentiable at $\mathfrak{u} \in \mathcal{A}^J$; then the right-hand side of (3.6) is locally Lipschitz as a mapping of the finite-dimensional space $\mathcal{A}^J$ into itself and standard ODE theory gives a local in time solution. Lemma 3.2 verifies the required differentiability. ∎

LEMMA 3.2. *The function* $E : \mathcal{A}^J \to \mathcal{L}(\mathcal{A}^J, \mathcal{A}^J)$ *is Fréchet differentiable with respect to* $\mathfrak{u} \in \mathcal{A}^J$.

Proof. To prove this, we write down the Fréchet partial derivative of each component of E with respect to $u^{(i)}$, applied in perturbation direction $v \in \mathcal{A}$; we use $\nabla G(u)$ to denote the Fréchet derivative of $G : K \to Y$ at point $u \in K$. Now note that

$$\begin{aligned}\left\langle v, \frac{\partial}{\partial u^{(i)}} D_{jk}(\mathfrak{u}) \right\rangle_K &= -\frac{1}{J}\langle \Gamma^{-1/2}(G(u^{(j)}) - y), \Gamma^{-1/2}\nabla G(u^{(i)})v\rangle_Y \\ &\quad + \mathbf{1}_{i=j}\langle \Gamma^{-1/2}\nabla G(u^{(j)})v, \Gamma^{-1/2}(G(u^{(k)}) - \bar{G})\rangle_Y \\ &\quad + \mathbf{1}_{i=k}\langle \Gamma^{-1/2}(G(u^{(j)}) - y), \Gamma^{-1/2}\nabla G(u^{(k)})v\rangle_Y.\end{aligned}$$

When G is C^1, $\nabla G(u^{(i)})$ is a bounded operator from K to Y, so the quantity above is bounded. Next, we define

$$P_{jk}(\mathfrak{u}) := \langle u^{(j)}, u^{(k)} - \bar{u}\rangle_K;$$

this is finite because it is bounded above by $\|u^{(j)}\|_K \|u^{(k)} - \bar{u}\|_K$ using the Cauchy–Schwarz inequality. Then

$$\left\langle v, \frac{\partial}{\partial u^{(i)}} P_{jk}(\mathfrak{u}) \right\rangle_K = -\frac{1}{J}\langle u^{(j)}, v\rangle_K + \mathbf{1}_{i=j}\langle v, u^{(k)} - \bar{u}\rangle_K + \mathbf{1}_{i=k}\langle u^{(j)}, v\rangle_K.$$

It is straightforward to verify that this is bounded for $v \in \mathcal{A}$. Since E is formed by summing D and P, the proof is complete. ∎

3.3. Ensemble collapse for TEKI flow. From Theorem 3.1 we know that the vector space $\mathcal{A}$ is invariant for the TEKI flow. Furthermore, when restricted to $\mathcal{A}$, C_0 is positive definite, so $\|\cdot\|_K = \|C_0^{-1/2}\cdot\|_X$ and $\|\cdot\|_X$ are equivalent norms on the vector space $\mathcal{A}$. In particular, the following constants are well defined and strictly positive:

$$\lambda_m(\mathcal{A}) := \inf_{v\in\mathcal{A},\|v\|_X^2=1} \|v\|_K^2, \quad \lambda_M(\mathcal{A}) := \sup_{v\in\mathcal{A},\|v\|_X^2=1} \|v\|_K^2. \tag{3.11}$$

Note that $\lambda_m(\mathcal{A})$ and $\lambda_M(\mathcal{A})$ do depend on $\mathcal{A}$, which is defined through the initial choice of ensemble members.

The empirical covariance $C(\mathfrak{u}(t))$ can also be viewed as a matrix in the finite-dimensional linear space $\mathcal{A}$. The following theorem demonstrates that its operator norm can be bounded from above uniformly in time, and establishes asymptotic in time collapse of the ensemble, provided that the solution exists for all time.

THEOREM 3.3. *For the TEKI flow defined by* (3.5) *the following upper bound holds while a solution exists:*

$$\|C(\mathfrak{u}(t))\|_X \le \frac{1}{\|C(\mathfrak{u}(0))\|_X^{-1} + 2\lambda_m(\mathcal{A})t}. \tag{3.12}$$

Here $\|C(\mathfrak{u}(t))\|_X$ *is the operator norm of* $C(\mathfrak{u}(t))$ *on* $(\mathcal{A}, \|\cdot\|_X)$ *and* $\lambda_m(\mathcal{A})$ *is defined in* (3.11).

Proof. Recall the dynamical system for $u^{(j)}(t)$:

$$\frac{d}{dt}u^{(j)} = -\frac{1}{J}\sum_{k=1}^{J}\Big(D_{jk}(\mathfrak{u}) + \langle u^{(j)}, u^{(k)} - \bar{u}\rangle_K\Big)(u^{(k)} - \bar{u}).$$

Averaging over j, we have the ODE for $\bar{u}(t)$. Taking the difference, we have

$$\begin{aligned}\frac{d}{dt}(u^{(j)} - \bar{u}) = &-\frac{1}{J}\sum_{k=1}^{J}\langle\Gamma^{-1/2}(G(u^{(j)}) - \bar{G}), \Gamma^{-1/2}(G(u^{(k)}) - \bar{G})\rangle_Y(u^{(k)} - \bar{u})\\ &- \frac{1}{J}\sum_{k=1}^{J}\langle u^{(j)} - \bar{u}, u^{(k)} - \bar{u}\rangle_K(u^{(k)} - \bar{u}).\end{aligned}$$

Then, because $C(\mathfrak{u}(t)) = \frac{1}{J}\sum_{j=1}^{J}(u^{(j)}(t) - \bar{u}(t)) \otimes (u^{(j)}(t) - \bar{u}(t))$, we find that

$$\begin{aligned}\frac{dC(\mathfrak{u}(t))}{dt} = &-\frac{2}{J^2}\sum_{j,k=1}^{J}\langle u^{(j)} - \bar{u}, u^{(k)} - \bar{u}\rangle_K(u^{(k)} - \bar{u}) \otimes (u^{(j)} - \bar{u})\\ &- \frac{2}{J^2}\sum_{j,k=1}^{J}\langle\Gamma^{-1/2}(G(u^{(j)}) - \bar{G}), \Gamma^{-1/2}(G(u^{(k)}) - \bar{G})\rangle_Y(u^{(k)} - \bar{u}) \otimes (u^{(j)} - \bar{u}).\end{aligned}$$

Now we consider projecting the ODE above on a fixed $v \in X$. Denote

$$v_k(t) = \langle v, u^{(k)}(t)\rangle_X, \quad \bar{v}(t) = \langle v, \bar{u}(t)\rangle_X.$$

Note that

$$\langle v, (u^{(k)} - \bar{u}) \otimes (u^{(j)} - \bar{u})v\rangle_X = \langle v, u^{(k)} - \bar{u}\rangle_X\langle v, u^{(j)} - \bar{u}\rangle_X = (v^{(k)} - \bar{v})(v^{(j)} - \bar{v}).$$

The projection of $\frac{dC(\mathfrak{u}(t))}{dt}$ on v is given by

$$
\begin{aligned}
&J^2\left\langle v, \frac{d}{dt}C(\mathfrak{u}(t))v\right\rangle_X \\
&= -2\sum_{j,k=1}^{J}\langle u^{(j)}-\bar{u}, u^{(k)}-\bar{u}\rangle_K (v_k-\bar{v})\cdot(v_j-\bar{v}) \\
&\quad -2\sum_{j,k=1}^{J}\langle \Gamma^{-1/2}(G(u^{(j)})-\bar{G}), \Gamma^{-1/2}(G(u^{(k)})-\bar{G})\rangle_Y (v_k-\bar{v})\cdot(v_j-\bar{v}) \\
&= -2\left\|\sum_{j=1}^{J}(v_j-\bar{v})(u^{(j)}-\bar{u})\right\|_K^2 - 2\left\|\Gamma^{-1/2}\sum_{j=1}^{J}(v_j-\bar{v})(G(u^{(j)})-\bar{G})\right\|_Y^2. \qquad (3.13)
\end{aligned}
$$

Note that

$$
\frac{1}{J}\sum_{j=1}^{J}(v_j-\bar{v})(u^{(j)}-\bar{u}) = \frac{1}{J}\sum_{j=1}^{J}\langle v, u^{(j)}-\bar{u}\rangle_X (u^{(j)}-\bar{u}) = C(\mathfrak{u})v,
$$

so if $v\in\mathcal{A}$, then

$$
\left\langle v, \frac{d}{dt}C(\mathfrak{u}(t))v\right\rangle_X \le -2\|C(\mathfrak{u})v\|_K^2 \le -2\lambda_m(\mathcal{A})\|C(\mathfrak{u}(t))v\|_X^2.
$$

Here we used that for all $v\in\mathcal{A}$,

$$
C(\mathfrak{u}(t))v = \frac{1}{J}\sum_{j=1}^{J}\langle v, (u^{(j)}-\bar{u})\rangle_X (u^{(j)}-\bar{u}) \in \mathcal{A}.
$$

Consider $C(\mathfrak{u}(t))$ as a matrix in $(\mathcal{A}, \|\cdot\|_X)$, and let $w(t)$ be the unit-norm eigenvector with maximum eigenvalue. Since

$$
0 = \frac{d}{dt}\|w(t)\|_X^2 = 2\left\langle w(t), \frac{d}{dt}w(t)\right\rangle_X,
$$

it follows that

$$
\begin{aligned}
\frac{d}{dt}\|C(\mathfrak{u}(t))\|_X &= \frac{d}{dt}\langle w(t), C(\mathfrak{u}(t))w(t)\rangle_X \\
&= \left\langle w, \frac{d}{dt}C(\mathfrak{u})w\right\rangle_X + 2\left\langle \frac{d}{dt}w(t), C(\mathfrak{u}(t))w(t)\right\rangle_X \\
&= \left\langle w, \frac{d}{dt}C(\mathfrak{u})w\right\rangle_X + 2\|C(\mathfrak{u})\|_K\left\langle \frac{d}{dt}w(t), w(t)\right\rangle_X \\
&\le -2\lambda_m(\mathcal{A})\|C(\mathfrak{u}(t))w(t)\|_X^2 = -2\lambda_m(\mathcal{A})\|C(\mathfrak{u}(t))\|_X^2.
\end{aligned}
$$

So

$$
\frac{d}{dt}\|C(\mathfrak{u}(t))\|_X^{-1} = -\|C(\mathfrak{u}(t))\|_X^{-2}\frac{d}{dt}\|C(\mathfrak{u}(t))\|_X \ge 2\lambda_m(\mathcal{A}),
$$

and hence we have our claim. ∎

Remark 3.3. The bound in the preceding theorem shows that the TEKI ensemble collapses, even in the case of nonlinear G; previous collapse results for EKI concern only the linear setting. The rate of collapse for each ensemble member is $\mathcal{O}(\frac{1}{\sqrt{t}})$. In classical Kalman filter theory, upper bounds for the covariance matrix can be obtained through an observability condition. In the TEKI algorithm the inclusion of a (prior) observation u in $F(u)$ enforces observability of the system. This provides the intuition for the upper bound (3.12) we prove for the TEKI covariance.

We conclude this subsection with a lemma and corollary which dig a little deeper into the properties of the solution ensemble, within the invariant subspace $\mathcal{A}$.

LEMMA 3.4. *For any $u^\perp \in X$, if $\langle u^\perp, u^{(j)}(0) - \bar{u}(0)\rangle_X = 0$ for all $j = 1, \dots, J$, then the TEKI flow will not change along the direction of $u^\perp$ while the solution exists:*

$$\langle u^\perp, u^{(j)}(t)\rangle_X = \langle u^\perp, \bar{u}(0)\rangle_X.$$

In particular $C(\mathfrak{u}(t))u^\perp \equiv \mathbf{0}$.

Proof. First of all, recall that (3.13) holds for all $v \in X$. We let $v = u^\perp$, which leads to

$$0 \le \langle u^\perp, C(\mathfrak{u}(t))u^\perp\rangle_X \le \langle u^\perp, C(\mathfrak{u}(0))u^\perp\rangle_X = \frac{1}{J}\sum_{j=1}^{J}\langle u^\perp, u^{(j)}(0) - \bar{u}(0)\rangle_X^2 = 0.$$

Then from

$$\langle u^\perp, C(\mathfrak{u}(t))u^\perp\rangle_X = \frac{1}{J}\sum_{j=1}^{J}\langle u^\perp, u^{(j)}(t) - \bar{u}(t)\rangle_X^2$$

we find that $\langle u^\perp, u^{(j)}(t) - \bar{u}(t)\rangle_X = 0$. Next, we note that

$$\begin{aligned}&\frac{d}{dt}\langle u^\perp, u^{(j)}(t)\rangle_X\\ &= -\frac{1}{J}\sum_{k=1}^{J}\Big(D_{jk}(\mathfrak{u}) + \langle C_0^{-1/2}u^{(j)}, C_0^{-1/2}(u^{(k)} - \bar{u})\rangle_X\Big)\langle u^\perp, u^{(k)} - \bar{u}\rangle_X = 0.\end{aligned}$$

So $\langle u^\perp, u^{(j)}(t)\rangle_X = \langle u^\perp, u^{(j)}(0)\rangle_X = \langle u^\perp, \bar{u}(0)\rangle_X$. Last, for any fixed v,

$$\langle v, C(\mathfrak{u}(t))u^\perp\rangle_X = \frac{1}{J}\sum_{j=1}^{J}\langle v, u^{(j)}(t) - \bar{u}(t)\rangle_X\langle u^\perp, u^{(j)}(t) - \bar{u}(t)\rangle_X = 0.$$

So $C(\mathfrak{u}(t))u^\perp \equiv \mathbf{0}$. ∎

Lemma 3.4 suggests that we define the following subspace $\mathcal{B} \subseteq \mathcal{A}$:

$$\mathcal{B} := \operatorname{span}\{u^{(j)}(0) - \bar{u}(0),\ j = 1, \dots, J\}.$$

Let $P_{\mathcal{B}}$ be the orthogonal projection onto $\mathcal{B}$ with respect to $\|\cdot\|_X$, and let

$$u_0^\perp := \bar{u}(0) - P_{\mathcal{B}}\bar{u}(0). \tag{3.14}$$

For notational simplicity we write $v \perp \mathcal{B}$ if $\langle v, u\rangle_X = 0$ for all $u \in \mathcal{B}$. Then $u_0^\perp \perp \mathcal{B}$, and $u^{(j)}(0) - u_0^\perp \in \mathcal{B}$ for all j. By Lemma 3.4, we know, for any $v \perp \mathcal{B}$, that

$$\langle v, u^{(j)}(t)\rangle_X = \langle v, u^{(j)}(0)\rangle_X = \langle v, u_0^\perp\rangle_X \Leftrightarrow \langle v, u^{(j)}(t) - u_0^\perp\rangle_X = 0.$$

In other words we further improve the results in Theorem 3.1 to the following corollary.

COROLLARY 3.5. *The TEKI flow stays in the affine space $u_0^\perp + \mathcal{B}$, that is,*

$$u^{(j)}(t) - u_0^\perp \in \mathcal{B} \quad \textit{while the solution exists.}$$

3.4. A priori bounds on TEKI flow. In many inverse problems prior information is available in terms of rough upper estimates on $\|u\|_K^2$, where K is an appropriately chosen Banach space. Classically, Tikhonov regularization is used to achieve such bounds, and in this subsection we show how similar bounds may be imposed through the TEKI flow approach. In the study of the EnKF for state estimation some general conditions that guarantee boundedness of the solutions are investigated [25, 43]. However, in general, EnKF-based state estimation can exhibit a catastrophic growth phenomenon [24]. For inverse problems, and TEKI in particular, the situation is more favorable. We study the linear setting first and then the nonlinear case. Recall the definition (3.9) of $\mathcal{I}_{\text{linear}}$.

PROPOSITION 3.6. *If the observation operator G is linear and bounded, then the TEKI flow* (3.6) *has a solution $u \in C([0,\infty), \mathcal{A})$ and, for all $t \geq 0$,*

$$\|u^{(j)}(t)\|_K^2 \leq 2\mathcal{I}_{\text{linear}}(u^{(j)}(0); y).$$

Proof. Simply note that in the linear case, the TEKI flow can be written as a gradient flow in the form (3.10), so that

$$\frac{d}{dt}\mathcal{I}_{\text{linear}}(u^{(j)}(t); y) = -\langle \nabla_u \mathcal{I}_{\text{linear}}(u^{(j)}(t); y), C(\mathfrak{u})\nabla_u \mathcal{I}_{\text{linear}}(u^{(j)}(t); y)\rangle_K \leq 0.$$

Therefore, $\mathcal{I}_{\text{linear}}(u^{(j)}(t); y) \leq \mathcal{I}_{\text{linear}}(u^{(j)}(0); y)$. This implies that

$$\frac{1}{2}\|u^{(j)}(t)\|_K^2 = R(u^{(j)}(t)) \leq \mathcal{I}_{\text{linear}}(u^{(j)}(t); y) \leq \mathcal{I}_{\text{linear}}(u^{(j)}(0); y).$$

As the solution is bounded, it cannot blow up and hence, because the dynamics are finite-dimensional, exists for all time. ∎

It is difficult to show that TEKI flow is bounded for a general, nonlinear observation operator. However, by modifying the observation operator outside a bounded set, it is possible to obtain bounds on the TEKI flow, using the regularization term to provide the needed control. Modification of the observation operator for $\|u\|_K$ sufficiently large is quite natural if one has a prior knowledge of where the solutions of the inverse problem lie. We seek a solution satisfying $\|u\|_K \leq M$ for some known constant M and define

$$\widetilde{G}(u) = \phi_M(\|u\|_K)G(u), \tag{3.15}$$

where $\phi_M(x)$ is a smooth transition function satisfying $\phi_M(x) = 1$ if $x < M$ and $\phi_M(x) = 0$ if $x > M + 1$. Using $\widetilde{G}(u)$ instead of G is natural in situations where we seek solutions satisfying $\|u\|_K \leq M$. To understand this setting, we work in the remainder of this section under the following assumption.

ASSUMPTION 3.7. *There is a constant M, so that $G(u) = \mathbf{0}$ if $\|u\|_K > M + 1$.*

For our results below to hold, $G(u)$ can take any fixed constant value when $\|u\|_K > M + 1$. We choose $\mathbf{0}$ to be concrete. One can then apply simple modifications such as (3.15). The advantage of Assumption 3.7 is that, when the ensemble is outside $\{u : \|u\|_K > M + 1\}$, the data misfit has no effect, and the TEKI flow is forced only by the gradient of the regularization term. This gradient controls TEKI if it is outside the ball $\{u : \|u\|_K > M + 1\}$.

PROPOSITION 3.8. *Let Assumption* 3.7 *hold. Then for any fixed* T *the TEKI flow has unique solution* $u \in C([0,\infty),\mathcal{A})$ *satisfying, for every ensemble member* j*,*

$$\sup_{t\geq T}\|u^{(j)}(t)\|_K \leq \max\left\{\|u^{(j)}(T)\|_K, M+\sqrt{\frac{2\lambda_M(\mathcal{A})J}{\lambda_m(\mathcal{A})T}}+1\right\}.$$

The constants $\lambda_m(\mathcal{A})$ *and* $\lambda_M(\mathcal{A})$ *are given by* (3.11)*.*

Proof. By Theorem 3.3, we deduce that, assuming a solution exists for all time,

$$\sup_{t\geq T}\|C(\mathfrak{u}(t))\|_X < \frac{1}{2\lambda_m(\mathcal{A})T}.$$

Note that, for any j,

$$\frac{1}{J}\|u^{(j)}-\bar{u}\|_X^4 \leq \langle u^{(j)}-\bar{u}, C(\mathfrak{u})(u^{(j)}-\bar{u})\rangle_X \leq \|C(\mathfrak{u})\|_X\|u^{(j)}-\bar{u}\|_X^2.$$

As a consequence, assuming that a solution exists for all time, then every ensemble member j satisfies

$$\sup_{t\geq T}\|u^{(j)}(t)-\bar{u}(t)\|_X^2 \leq J\sup_{t\geq T}\|C(\mathfrak{u}(t))\|_X < \frac{J}{2\lambda_m(\mathcal{A})T}.$$

Therefore, again assuming that a solution exists for all time, every ensemble member $u^{(j)}$ satisfies

$$\sup_{t\geq T}\|u^{(j)}(t)-\bar{u}(t)\|_K^2 \leq \lambda_M(\mathcal{A})\sup_{t\geq T}\|u^{(j)}(t)-\bar{u}(t)\|_X^2 \leq \frac{\lambda_M(\mathcal{A})J}{2\lambda_m(\mathcal{A})T}. \tag{3.16}$$

Now assume that for some ensemble member $u^{(k)}$ and some time $t \geq T$ we have

$$\|u^{(k)}(t)\|_K > M+2\sqrt{\frac{\lambda_M(\mathcal{A})J}{2\lambda_m(\mathcal{A})T}}+1. \tag{3.17}$$

It follows from (3.16) with $j=k$ that, for all $t \geq T$,

$$\|u^{(k)}(t)\|_K - \|\bar{u}(t)\|_K \leq \sqrt{\frac{\lambda_M(\mathcal{A})J}{2\lambda_m(\mathcal{A})T}}$$

and hence that

$$\|\bar{u}(t)\|_K \geq M+\sqrt{\frac{\lambda_M(\mathcal{A})J}{2\lambda_m(\mathcal{A})T}}+1.$$

Now from (3.16) with any index j we deduce that, for all $t \geq T$,

$$\|\bar{u}(t)\|_K - \|u^{(j)}(t)\|_K \leq \sqrt{\frac{\lambda_M(\mathcal{A})J}{2\lambda_m(\mathcal{A})T}}$$

and hence that, for any ensemble member $u^{(j)}$,

$$\|u^{(j)}(t)\|_K \geq M+1.$$

It follows that if (3.17) holds, then

$$D_{k\ell}(\mathfrak{u}) = \langle \Gamma^{-1/2}(G(u^{(k)}) - y), \Gamma^{-1/2}(G(u^{(\ell)}) - \bar{G})\rangle_Y = 0.$$

Then

$$\frac{d}{dt}u^{(k)}(t) = -C(\mathfrak{u})C_0^{-1}u^{(k)} \Rightarrow \frac{d}{dt}\|u^{(k)}(t)\|_K^2 = -2\langle C_0^{-1}u^{(k)}, C(\mathfrak{u})C_0^{-1}u^{(k)}\rangle_X \leq 0.$$

It follows that, for $t \geq T$, the function $t \mapsto \|u^{(k)}(t)\|_K$ is nonincreasing whenever it is larger than $M + \sqrt{\frac{2\lambda_M(\mathcal{A})J}{\lambda_m(\mathcal{A})T}} + 1$. This demonstrates the desired upper bound on the solution which, in turn, proves global existence of a solution to (3.5). ∎

3.5. Long-time analysis for TEKI flow: The linear setting. Theorem 3.3 shows that the TEKI ensemble collapses as time evolves. As the collapse is approached, it is natural to use a linear approximation to understand the TEKI flow. This motivates the analysis in this subsection where we consider the linear setting $G(u) = Au$ and study the asymptotic behavior of the TEKI flow. The following theorem will be developed only in the finite-dimensional setting; development in various infinite-dimensional settings, guided by specific linear inverse problems of applied interest, would constitute valuable future research.

ASSUMPTION 3.9. *Both X and Y are finite-dimensional spaces, and the matrix C_0 is strictly positive definite on X.*

From Corollary 3.5 we know the TEKI flow is restricted to the affine subspace $u_0^\perp + \mathcal{B} \subset K$. Given this constraint, it is natural to expect the long-time limit point of $u^{(j)}(t)$ to be of form $u_0^\perp + u_\mathcal{B}^\dagger$, where

$$u_\mathcal{B}^\dagger = \arg\min_{u\in\mathcal{B}}\Big\{\|C_0^{-1/2}(u + u_0^\perp)\|_X^2 + \|\Gamma^{-1/2}(A(u + u_0^\perp) - y)\|_Y^2\Big\}.$$

Then the Karush–Kuhn–Tucker (KKT) condition yields that

$$(C_0^{-1} + A^*\Gamma^{-1}A)(u_\mathcal{B}^\dagger + u_0^\perp) - A^*\Gamma^{-1}y =: v^\dagger \perp \mathcal{B}.$$

Here A^* is the adjoint of $A : (X, \|\cdot\|_X) \to (Y, \|\cdot\|_Y)$.

Note that $\Omega := C_0^{-1} + A^*\Gamma^{-1}A$ is the posterior precision matrix of the Bayesian inverse problem associated to inverting A subject to additive Gaussian noise $N(0, \Gamma)$ and prior $N(0, C_0)$ on u. Since we often consider elements in the subspace $\mathcal{B}$, we also denote the restriction of Ω in $\mathcal{B}$ as Ω_B. Note that $0 < \langle u, \Omega_\mathcal{B} u\rangle_X < \infty$ for all nontrivial $u \in \mathcal{B}$, and $\Omega_\mathcal{B}$ is positive definite on $\mathcal{B}$, while $\Omega_\mathcal{B}^{-1}$ and $\Omega_\mathcal{B}^{1/2}$ are both well defined.

THEOREM 3.10. *Let Assumption 3.9 hold, and assume further that $G(u) = Au$, where A is a bounded linear map. Then the TEKI flow exists for all $t > 0$ and the solution converges to $u_0^\perp + u_\mathcal{B}^\dagger$ with rate of $O(\frac{1}{\sqrt{t}})$. In particular, $e^{(j)}(t) = u^{(j)}(t) - u_0^\perp - u_\mathcal{B}^\dagger$ is bounded by*

$$\|e^{(j)}(t)\|_Z^2 \leq \frac{m_0}{1 + 2m_0 t}\|e^{(j)}(0)\|_Z^2.$$

Here the constant m_0 is given by

$$m_0 := \min_{u\in\mathcal{B}, \|u\|_X = 1} \langle \Omega_\mathcal{B}^{1/2}u, C(\mathfrak{u}(0))\Omega_\mathcal{B}^{1/2}u\rangle_X.$$

Furthermore, $\|\cdot\|_Z$ *is the norm on* $\mathcal{B}$ *given by*

$$\|u\|_Z^2 = \|u\|_K^2 + \|\Gamma^{-1/2}Au\|_Y^2 = \langle \Omega_{\mathcal{B}} u, u\rangle_X.$$

Before proving the theorem, we discuss how the constraint $u_{\mathcal{B}}^\dagger \in \mathcal{B}$ changes the solution in relation to the unconstrained optimization. For that purpose, consider the unconstrained problem

$$u^\dagger = \arg\min_{u\in K}\Big\{\|u + u_0^\perp\|_K^2 + \|\Gamma^{-1/2}(A(u+u_0^\perp) - y)\|_Y^2\Big\}.$$

(This corresponds to finding the maximum a posteriori estimator for the Bayesian inverse problem referred to above.) The KKT condition leads to

$$\Omega(u^\dagger + u_0^\perp) = A^*\Gamma^{-1}y.$$

Note that $u^\dagger$ is in the space K, while $u_{\mathcal{B}}^\dagger$ is in the subspace $\mathcal{B}$. It is natural to try and understand the relationship between $u^\dagger$ and $u_{\mathcal{B}}^\dagger$ since this sheds light on the optimal choice of $\mathcal{B}$ and hence of the initial ensembles. To this end, we have the following proposition.

PROPOSITION 3.11. *Under the same conditions as Theorem* 3.10*, let* $P_{\mathcal{B}}$ *be the orthogonal projection from* K *to* $\mathcal{B}$ *with respect to the inner product* $\langle\cdot\rangle_X$*, and let* $P_\perp = \mathbf{I} - P_{\mathcal{B}}$*. Then* $u_{\mathcal{B}}^\dagger$ *can be written as*

$$u_{\mathcal{B}}^\dagger = P_{\mathcal{B}}u^\dagger + \Omega_{\mathcal{B}}^{-1}P_{\mathcal{B}}\Omega P_\perp u^\dagger.$$

In particular, if $\mathcal{B}$ *and its orthogonal complement have no correlation through* Ω *(i.e.,* $\langle u, \Omega v\rangle_X = 0$ *for all* $u \in \mathcal{B}$ *and* $v \perp \mathcal{B}$*), then* $u_{\mathcal{B}}^\dagger = P_{\mathcal{B}}u^\dagger$*.*

Proof. Recall the KKT conditions

$$\Omega(u^\dagger + u_0^\perp) = A^*\Gamma^{-1}y, \quad \Omega(u_{\mathcal{B}}^\dagger + u_0^\perp) = A^*\Gamma^{-1}y + v^\dagger,$$

where $v^\dagger \perp \mathcal{B}$. They lead to

$$\Omega u_{\mathcal{B}}^\dagger = \Omega u^\dagger + v^\dagger = \Omega P_{\mathcal{B}}u^\dagger + \Omega P_\perp u^\dagger + v^\dagger.$$

Projecting this equation into $\mathcal{B}$, we find that

$$P_{\mathcal{B}}\Omega P_{\mathcal{B}}u_{\mathcal{B}}^\dagger = P_{\mathcal{B}}\Omega P_{\mathcal{B}}u^\dagger + P_{\mathcal{B}}\Omega P_\perp u^\dagger.$$

Note that for any $v_1, v_2 \in \mathcal{B}$, $\langle v_1, \Omega_{\mathcal{B}}v_2\rangle_X = \langle v_1, P_{\mathcal{B}}\Omega P_{\mathcal{B}}v_2\rangle_X$, so $\Omega_{\mathcal{B}}v_2 = P_{\mathcal{B}}\Omega P_{\mathcal{B}}v_2$. Therefore, we have

$$\Omega_{\mathcal{B}}(u_{\mathcal{B}}^\dagger - P_{\mathcal{B}}u^\dagger) = P_{\mathcal{B}}\Omega P_\perp u^\dagger.$$

Finally, note that $\Omega_{\mathcal{B}}$ is positive definite within $\mathcal{B}$ and hence invertible within $\mathcal{B}$. Applying $\Omega_{\mathcal{B}}^{-1}$ on both sides, we have our claim. ∎

Proof of Theorem 3.10. We investigate the dynamics of $e^{(j)}(t) = u^{(j)}(t) - u_0^{\perp} - u_{\mathcal{B}}^{\dagger} \in \mathcal{B}$. Note that

$$\begin{aligned}
\frac{d}{dt}e^{(j)}(t) &= -\frac{1}{J}\sum_{k=1}^{J}\Big(\langle \Gamma^{-1}(Au^{(j)} - y), A(u^{(k)} - \bar{u})\rangle_Y + \langle u^{(j)}, u^{(k)} - \bar{u}\rangle_K\Big)(u^{(k)} - \bar{u}) \\
&= -\frac{1}{J}\sum_{k=1}^{J}\langle A^*\Gamma^{-1}(Au^{(j)} - y), u^{(k)} - \bar{u}\rangle_X (u^{(k)} - \bar{u}) \\
&\quad - \frac{1}{J}\sum_{k=1}^{J}\langle u^{(j)}, C_0^{-1}(u^{(k)} - \bar{u})\rangle_X (u^{(k)} - \bar{u}) \\
&= -C(\mathfrak{u})A^*\Gamma^{-1}(Au^{(j)} - y) - C(\mathfrak{u})C_0^{-1}u^{(j)} \\
&= -C(\mathfrak{u})(A^*\Gamma^{-1}A + C_0^{-1})u^{(j)} + C(\mathfrak{u})A^*\Gamma^{-1}y \\
&= -C(\mathfrak{u})(A^*\Gamma^{-1}A + C_0^{-1})u^{(j)} + C(\mathfrak{u})(C_0^{-1} + A^*\Gamma^{-1}A)\Big(u_{\mathcal{B}}^{\dagger} + u_0^{\perp} - v^{\dagger}\Big) \\
&= -C(\mathfrak{u})(A^*\Gamma^{-1}A + C_0^{-1})e^{(j)}(t) - C(\mathfrak{u})v^{\dagger}.
\end{aligned}$$

But Lemma 3.4 shows that $C(\mathfrak{u})v^{\dagger} = \mathbf{0}$, so we have established that

$$\frac{d}{dt}e^{(j)}(t) = -C(\mathfrak{u}(t))\Omega e^{(j)}(t).$$

Since we know that $e^{(j)}(t) \in \mathcal{B}$, $C(\mathfrak{u}(t))w = \mathbf{0}$ for any $w \perp \mathcal{B}$, the equation above can be written as

$$\frac{d}{dt}e^{(j)}(t) = -C(\mathfrak{u}(t))\Omega_{\mathcal{B}}e^{(j)}(t).$$

This leads to

$$\begin{aligned}
\frac{1}{2}\frac{d}{dt}\|e^{(j)}(t)\|_Z^2 &= -\langle \Omega_{\mathcal{B}}e^{(j)}, C(\mathfrak{u})\Omega_{\mathcal{B}}e^{(j)}\rangle_X \\
&= -\langle \Omega_{\mathcal{B}}^{\frac{1}{2}}e^{(j)}, D(\mathfrak{u})\Omega_{\mathcal{B}}^{\frac{1}{2}}e^{(j)}\rangle_X,
\end{aligned}$$

where $D(\mathfrak{u}) = \Omega_{\mathcal{B}}^{\frac{1}{2}}C(\mathfrak{u})\Omega_{\mathcal{B}}^{\frac{1}{2}}$ on $\mathcal{B}$. Lemma 3.12 below shows that for any $v \in \mathcal{B}$ with $\|v\|_X = 1$, and m_0 as defined above,

$$\langle v, D(\mathfrak{u})v\rangle_X \geq \frac{1}{m_0^{-1} + 2t}.$$

Therefore,

$$\frac{d}{dt}\|e^{(j)}(t)\|_Z^2 \leq -\frac{2}{m_0^{-1} + 2t}\|\Omega_{\mathcal{B}}^{1/2}e^{(j)}(t)\|_K^2 = -\frac{2}{m_0^{-1} + 2t}\|e^{(j)}(t)\|_Z^2.$$

This leads to

$$\frac{d}{dt}\log\|e^{(j)}(t)\|_Z^2 \leq -\frac{2}{m_0^{-1} + 2t} \Rightarrow \|e^{(j)}(t)\|_Z^2 \leq \frac{1}{1 + 2m_0 t}\|e^{(j)}(0)\|_Z^2. \qquad \square$$

LEMMA 3.12. *Let the same conditions as in Theorem* 3.10 *hold, and define* $D(\mathfrak{u}) = \Omega_{\mathcal{B}}^{1/2} C(\mathfrak{u}) \Omega_{\mathcal{B}}^{1/2}$. *Then, given any* $v \in \mathcal{B}, \|v\|_Z = 1$,

$$\langle \Omega_{\mathcal{B}}^{1/2} v, D(\mathfrak{u}(t)) \Omega_{\mathcal{B}}^{1/2} v \rangle_X \geq \frac{1}{m_0^{-1} + 2t}.$$

Proof. Recall (3.13) and set $G(u) = Au$ to obtain

$$\begin{aligned}
\left\langle v, \frac{d}{dt} C(\mathfrak{u}(t)) v \right\rangle_X &= -\frac{2}{J^2} \left\| \sum_{j=1}^{J} (v_j - \bar{v})(u^{(j)} - \bar{u}) \right\|_K^2 \\
&\quad - \frac{2}{J^2} \left\| \Gamma^{-1/2} A \sum_{j=1}^{J} (v_j - \bar{v})(u^{(j)} - \bar{u}) \right\|_Y^2 \\
&= -2 \| C_0^{-1/2} C(\mathfrak{u}(t)) v \|_K^2 - 2 \| \Gamma^{-1/2} A C(\mathfrak{u}(t)) v \|_Y^2 \\
&= -2 \langle C(\mathfrak{u}(t)) v, (C_0^{-1} + A^* \Gamma^{-1} A) C(\mathfrak{u}(t)) v \rangle_X.
\end{aligned}$$

Since this is true for any v, we deduce that $C(\mathfrak{u}(t))$ as a matrix on $\mathcal{B}$ satisfies

$$\frac{d}{dt} C(\mathfrak{u}(t)) = -2C(\mathfrak{u}(t))(C_0^{-1} + A^* \Gamma^{-1} A) C(\mathfrak{u}(t)).$$

Recall that by Lemma 3.4, $C(\mathfrak{u}) v = \mathbf{0}$ for all $v \perp \mathcal{B}$. As a consequence, $C(\mathfrak{u}) = P_{\mathcal{B}} C(\mathfrak{u}) P_{\mathcal{B}}$, and therefore we can write

$$\frac{d}{dt} C(\mathfrak{u}(t)) = -2C(\mathfrak{u}(t)) \Omega_{\mathcal{B}} C(\mathfrak{u}(t)).$$

So, by the chain rule,

$$\frac{d}{dt} D(\mathfrak{u}(t)) = -2D(\mathfrak{u}(t))^2.$$

As a consequence, we find that each eigenvector v of $D(\mathfrak{u}(0))$ remains an eigenvector of $D(\mathfrak{u}(t))$, and its eigenvalue $\lambda = \lambda(t)$ solves the ODE

$$\frac{d}{dt} \lambda(t) = \frac{d}{dt} \langle v, D(\mathfrak{u}(t)) v \rangle_X = -\langle v, D(\mathfrak{u}(t))^2 v \rangle_X = -2\lambda^2.$$

The solution is given by $\lambda(t) = \frac{1}{\lambda(0)^{-1} + 2t}$. Taking $\lambda(0)$ to be the minimum eigenvalue of $D(\mathfrak{u}(0))$ gives our claim. ∎

4. Numerical experiments. In this section we describe numerical results comparing EKI with the regularized TEKI method. Our EKI and TEKI algorithms are based on time-discretizations of the continuum limit, rather than on the discrete algorithms stated in sections 1 and 2; we describe the adaptive time-steppers used in subsection 4.1. In subsection 4.2 we present the spectral discretization used to create prior samples and demonstrate how to introduce the additional regularization of prior samples required for the TEKI approach. Subsection 4.3 contains numerical experiments comparing EKI and TEKI. The inverse problem is to find the slowness function in an eikonal equation, given noisy travel time data. In subsection 4.4 we have also conducted numerical experiments with a porous medium equation. The inverse problem associated with it is to find the permeability, given noisy pressure measurements.

4.1. Temporal discretization. The specific ensemble Kalman algorithms that we use are found by applying the Euler discretization to the continuous time limit of each algorithm. Discretizing (3.5) with adaptive time-step h_n gives

$$u_{n+1}^{(j)} = u_n^{(j)} - \frac{h_n}{J}\sum_{k=1}^{J} E_{jk}(\mathfrak{u}_n)(u_n^{(k)} - \bar{u}_n) \tag{4.1a}$$

$$= u_n^{(j)} - \frac{h_n}{J}\sum_{k=1}^{J}\Big(D_{jk}(\mathfrak{u}_n) + \langle C_0^{-1}u_n^{(j)}, u_n^{(k)} - \bar{u}_n\rangle_X\Big)(u_n^{(k)} - \bar{u}_n). \tag{4.1b}$$

For the adaptive time-step we take, as implemented in [26],

$$h_n = \frac{h_0}{\|E(\mathfrak{u}_n)\|_F + \delta} \tag{4.2}$$

for some $h_0, \delta \ll 1$, where $\|\cdot\|_F$ denotes the Frobenius norm, and E is the matrix with entries E_{jk} (rather than its Kronecker form used earlier in (3.6)). The integration method for the EKI flow (3.1) is identical but with E_{jk} replaced by D_{jk}. Note that this adaptive time-step is motivated by an understanding of the stability restriction on the Euler method which arises for linear problems, here adapted to the state-dependence of D and E.

4.2. Spatial discretization. We consider all inverse problems on the two-dimensional spatial domain $\mathcal{D} = [0,1]^2$. We let $-\triangle$ denote the Laplacian on $\mathcal{D}$ subject to homogeneous Neumann boundary conditions. We then define

$$C_0 = \left(-\triangle + \tau^2\right)^{-\alpha},$$

where $\tau \in \mathbb{R}^+$ denotes the inverse lengthscale of the random field and $\alpha \in \mathbb{R}^+$ determines the regularity; specifically, draws from the random field are Hölder with exponent up to $\alpha - 1$ (since spatial dimension $d = 2$). From this we note that the eigenvalue problem

$$C_0\varphi_k = \lambda_k\varphi_k$$

has solutions for $\mathbb{Z} = \{0, 1, 2, \dots\}$:

$$\varphi_k(x) = \sqrt{2}\cos(k\pi x), \quad \lambda_k = \left(|k|^2\pi^2 + \tau^2\right)^{-\alpha}, \quad k \in \mathbb{Z}_+^2.$$

Here $X = L^2(\mathcal{D}, \mathbb{R})$ and the φ_k are orthonormal in X with respect to the standard inner product. Draws from the measure $N(0, C_0)$ are given by the Karhunen–Loève (KL) expansion

$$u = \sum_{k\in\mathbb{Z}_+^2} \sqrt{\lambda_k}\xi_k\varphi_k(x), \quad \xi_k \sim N(0,1) \quad \text{i.i.d.} \tag{4.3}$$

This random function will be almost surely in X and in $C(\mathcal{D}, \mathbb{R})$, provided that $\alpha > 1$ [41], and we therefore impose this condition.

Recall that for TEKI to be well defined we require an initial ensemble to lie in the Cameron–Martin space of the Gaussian measure $N(0, C_0)$. The draws in (4.3) do not satisfy this criterion; indeed, in infinite dimensions samples from Gaussian measure never live in the Cameron–Martin space. Instead, we consider an expansion in the form

$$v = \sum_{k\in\mathbb{Z}_+^2} \lambda_k^a\xi_k\varphi_k(x), \quad \xi_k \sim N(0,1) \quad \text{i.i.d.} \tag{4.4}$$

and determine a condition on a which ensures that such random functions lie in the domain of $C_0^{-\frac{1}{2}}$, the required Cameron–Martin space. We note that

$$\mathbb{E}\|v\|_K^2 = \mathbb{E}\|C_0^{-\frac{1}{2}}v\|_X^2 = \mathbb{E}\left\|\sum_{k\in\mathbb{Z}_+^2} \lambda_k^{a-\frac{1}{2}}\xi_k\varphi_k(x)\right\|_X^2 = \sum_{k\in\mathbb{Z}_+^2} \lambda_k^{2a-1}.$$

Since $\mathcal{D}$ is a two-dimensional domain, the eigenvalues of the Laplacian grow asymptotically like j if ordered on a one-dimensional lattice $\mathbb{Z}_+$ indexed by j. Thus it suffices to find a to ensure

$$\sum_{j\in\mathbb{Z}_+} j^{-\alpha(2a-1)} < \infty.$$

Hence we see that choosing $a > \frac{1}{2} + \frac{1}{2\alpha}$ will suffice. The initial ensemble for both EKI and TEKI is found by drawing functions v with a satisfying this inequality. The random function (4.4) is Hölder with exponent up to $2a\alpha - 1$.[1]

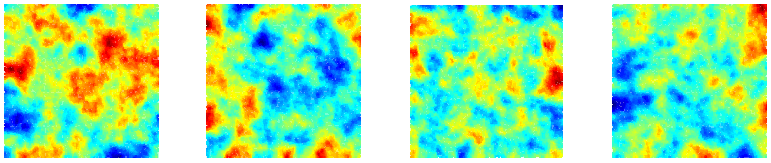

FIG. 1. *KL draws from the prior with $\alpha = 1$ and $a = 1$.*

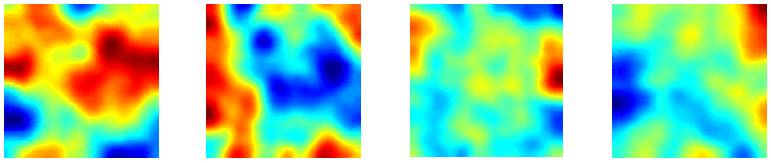

FIG. 2. *KL draws from the Cameron–Martin space of the prior with $\alpha = 2$ and $a = 1$.*

To illustrate the foregoing, we consider the Gaussian measure $N(0, C_0)$ which arises when $\alpha = 2$ and with inverse lengthscale $\tau = 15$. We study realizations from the KL expansion (4.3) and from the TEKI-regularized expansion (4.4) with $a = 1 > 3/4$, using common realizations of the random variables $\{\xi_k\}_{k\in\mathbb{Z}_+^2}$. Figure 1 shows four random draws from the KL expansion (4.3) and Figure 2 from (4.4). The required higher regularity of initial samples for the TEKI method is apparent. The functions in Figure 1 have Hölder exponent up to 1, while those in Figure 2 have Hölder exponent up to 3.

[1] For noninteger β we use the terminology that a function is Hölder with exponent β if the function is in $C^{\lfloor\beta\rfloor}$ and its $\lfloor\beta\rfloor$th derivatives are Hölder $\beta - \lfloor\beta\rfloor$. In the context of this paper integer β can be avoided because random Gaussian functions are always Hölder on an interval of exponents which is open from the right. See [10].

4.3. Inverse Eikonal equation. We test and compare EKI and TEKI on an inverse problem arising from the eikonal equation. This PDE arises in numerous scientific disciplines, in particular in seismic travel time tomography. Given a slowness or inverse velocity function $s(x) \in C^0(\bar{\mathcal{D}})$, characterizing the medium, and a source location $x_0 \in \mathcal{D}$, the forward eikonal equation is to solve for travel time $T(x) \in C^0(\bar{\mathcal{D}})$ satisfying

$$|\nabla T(x)| = s(x), \quad x \in \mathcal{D} \setminus \{x_0\}, \tag{4.5}$$

$$T(x_0) = 0, \tag{4.6}$$

$$\nabla T(x) \cdot \nu(x) \geq 0, \quad x \in \partial\mathcal{D}. \tag{4.7}$$

The forward solution $T(x)$ represents the shortest travel time from x_0 to a point in the domain $\mathcal{D}$. The Soner boundary condition (4.7) imposes wave propagation along the unit outward normal $\nu(x)$ on the boundary of the domain. For the slowness function $s(x)$ we assume the positivity $s(x) > 0$, which ensures well-posedness. The unique solution can be characterized via the minimization procedure found in [33].

The inverse problem is to determine the speed function s from measurements (linear mollified pointwise functionals $l_j(\cdot)$) of the travel time function T; for example, we might measure T at specific locations in the domain $\bar{\mathcal{D}}$. In order to ensure positivity of the speed function during inversion, we write $s = \exp(u)$ and invert for u rather than s. The data is assumed to take the form

$$y_j = l_j(T) + \eta_j, \quad j = 1, \ldots, J, \tag{4.8}$$

where the η_j are Gaussian noise, assumed independent, mean zero, and with covariance Γ. By defining $G_j(u) = l_j(T)$, we can rewrite (4.8) as the inverse problem

$$y = G(u) + \eta, \quad \eta \sim N(0, \Gamma). \tag{4.9}$$

Further details on the well-posedness of the forward and inverse eikonal equation can be found in Elliott, Deckelnick, and Styles [11].

The discretization of the forward model is based on a fast marching method [11, 40], employing a uniform mesh with spacing $h_* = 0.01$. On the left-hand boundary we choose five random source points with 64 equidistant pointwise measurements in the domain. For the inversion we choose $\Gamma = \gamma^2 I$, with $\gamma = 0.01$. We fix the ensemble size at $J = 100$ and the maximum number of iterations at 23. To define the adaptive time-stepping procedure, we take $h_0 = 0.02$ and $\delta = 0.05$.

Recall that the initial ensembles for EKI and TEKI, when chosen at random, differ in terms of regularity: TEKI draws lie in the Cameron–Martin space and hence are more regular than those for EKI. In order to thoroughly compare the methodologies we will consider three different truth functions $u^\dagger$, one each matching the regularities of the EKI and TEKI draws, respectively, and one with regularity lying between the regularities of the two EKI and TEKI initializations. The EKI draws in each of Cases 1, 2, and 3 are found by taking $\alpha = 2$ (and by definition $a = 0.5$) and the TEKI draws by taking $\alpha = 2$ and $a = 1$. The truth in each case is found by taking $\alpha = 2$, $a = 0.5$ (Case 1), $\alpha = 3.2$, $a = 0.5$ (Case 2), and $\alpha = 2$, $a = 1$ (Case 3). The resulting maximal Hölder exponents are shown in Table 1. We will also study the EKI and TEKI methods when initialized with the same initial ensemble, namely the KL eigenfunctions φ_k.

In addition to experiments where the initial ensembles are drawn at random from (4.3) (for EKI) and from (4.4) (for TEKI), we also consider experiments where the

TABLE 1
Maximal Hölder exponent for EKI and TEKI initial draws and truth $u^\dagger$.

Case	EKI	$u^\dagger$	TEKI
1	1	1	3
2	1	2.2	3
3	1	3	3

initial ensemble comprises the eigenfunctions

$$u^{(j)}(x) = \varphi_j(x), \quad j = 1, \dots, J, \tag{4.10}$$

and so it is the same for both EKI and TEKI. The first motivation for using the eigenfunctions is to facilitate a comparison between EKI and TEKI when they both use the same initial regularity, in contrast to the differing regularities in Table 1. The second motivation is that the choice of working with eigenfunctions, rather than random draws, has been show to guard against overfitting for EKI [21].

To assess the performance of both methods for each case, we consider analyzing this through two quantities, the relative error and the data misfit. These are defined, for EKI, as

$$\frac{|u^{(j)}_{\mathrm{EKI}} - u^\dagger|_2^2}{|u^\dagger|_2^2}, \quad \left|y - G(u^{(j)}_{\mathrm{EKI}})\right|_\Gamma,$$

and similarly for TEKI, where $|\cdot|_2$ denotes the standard Euclidean norm. When we evaluate these error and misfit measures, we will do so by employing the mean of the current ensemble. For the relative error we will plot this on a logarithmic scale. To see the effect of overfitting, we use the noise level $|\eta|_2 = |y - G(u^\dagger)|_2$ as a benchmark. While the observation model is nonlinear and hence does not directly follow Theorem 3.10, we plot the rate of $e(t) = \frac{1}{\sqrt{t}}$ for comparison. Throughout the experiments we show a progression through the $n = 23$ iterations, which will be represented through three subimages related to the (1st, 11th, 23rd) iterations, ordered from left to right. The first image, at step 1, is simply a single draw from the initial ensemble; the remaining two images show the mean of the ensemble at steps 11 and 23. We will plot the truth beside each set of progression images for ease of comparison. For the KL basis the image shown at step 1 is hence just one of the eigenfunctions φ_j. As mentioned, all of the numerics will be split into the three test cases as described in Table 1.

Remark 4.1. We note that for the purposes of all the results presented we set $\boldsymbol{\lambda} = 1$. We have conducted additional experiments for other values, including $\boldsymbol{\lambda} = 0.1, 10$, leading to no behavior majorly qualitatively different from that seen here. However, in general it will be of interest to learn the parameter λ, as is standard in the solution of ill-posed inverse problems [3, 13, 44]. We do not focus on this question here, however, as it distracts from the main message of the paper.

4.3.1. Case 1. Our first case corresponds to the first row of Table 1, as well as experiments in which both EKI and TEKI are initialized with the KL eigenfunctions. The truth and reconstructions are provided in Figure 3. We see no evidence of overfitting, and we notice that the TEKI solutions outperform the EKI solutions, and that the KL-initialized solutions are less accurate than those found from TEKI using random draws to initialize the ensemble; see Figure 4. Each row of Figure 3

demonstrates the progression of the method in each case. As the iteration progresses, we start to see differences in reconstruction for both EKI and TEKI. The regularity of the truth and the EKI initial ensemble match creating a superficial similarity in this case; however, the TEKI outperforms EKI despite this. When initializing with the KL basis, we notice a similar behavior for both TEKI and EKI. However, the added regularization for TEKI over EKI is manifest in a smaller error.

4.3.2. Case 2. Our second test case compares both methods when the regularity of the truth is between that of EKI and TEKI initial ensemble members. For this test case the truth and reconstructions are shown in Figure 5. The numerics for this test case show an ordering of the accuracy of the methods similar to that observed in Case 1. However, Figure 6 also demonstrates that the relative error of EKI with random draws starts to diverge. This is linked to the overfitting of the data since in this case the data misfit goes below the noise level. The results are similar to those obtained in [21] for EKI in the discrete form (1.3). This overfitting is demonstrated on the top row in Figure 5, which highlights the difficulty of reconstructing the truth within the linear span of the EKI initial ensemble.

For EKI and TEKI with a KL basis we see immediately that the divergence of the error does not occur here. Instead, the EKI algorithm performs relatively well, similarly to TEKI. However, the added regularization again leads to smaller errors in TEKI than in EKI. Interestingly, we also notice that there is little difference in TEKI for both the random draws and the KL basis. These results can be seen in the second and bottom rows of Figure 5. It is worth mentioning that, although Figure 6 shows that for TEKI with random draws the misfit reaches the noise level, running for further iterations does not result in overfitting (misfit falling below the noise level).

4.3.3. Case 3. Our third and final test case compares both methods, in a setting in which the regularity of the random draws for TEKI is the same as for the truth, shown in Figure 7. Figure 7 demonstrates almost identical outcomes as in Case 2, which show the progression of the iterations in the four different cases.

As the value of the regularity is higher compared to the previous case, we see the degeneracy of the EKI with random draws. This is highlighted in Figure 8, where we notice the same effect of the overfitting of the data as in Figure 6. This is similar to the top row of Figure 7, in that an overfitting phenomenon leads to a poor fitting of the truth as the iteration progresses.

All other methods, which include TEKI with random draws and both methods initialized with the KL basis, perform similarly. This can be attributed to the fact that all of their initial ensembles begin with a high regularity. As we observe from the third and bottom rows of Figure 7, the added regularization comes into play.

4.4. Darcy flow. Our final set of experiments will be to test the methodology on an inverse problem arising from Darcy flow. Given a domain $\mathcal{D} = [0,1]^2$ and real-valued permeability function $\kappa \in L^\infty(\mathcal{D})$ defined on $\mathcal{D}$, the forward model is concerned with determining a real-valued pressure (or hydraulic head) function $p \in H^1(\mathcal{D})$ on $\mathcal{D}$ from

$$-\nabla \cdot (\kappa \nabla p) = f, \quad x \in \mathcal{D},$$

with mixed boundary conditions

$$p(x_1, 0) = 100, \quad \frac{\partial p}{\partial x_1}(1, x_2) = 0, \quad -\kappa \frac{\partial p}{\partial x_1}(0, x_2) = 500, \quad \frac{\partial p}{\partial x_2}(x_1, 1) = 0$$

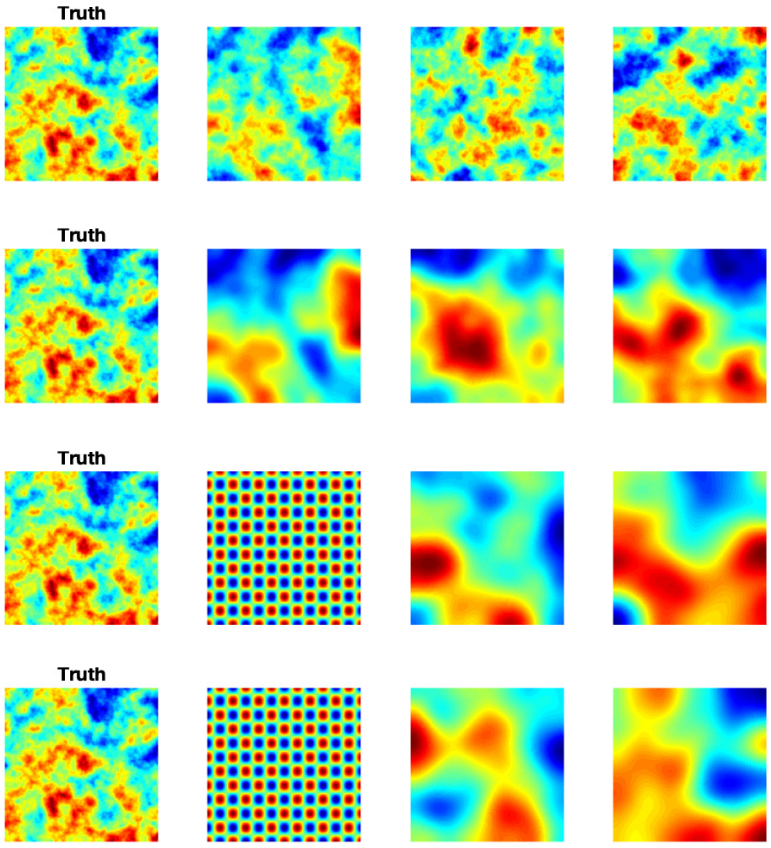

FIG. 3. *Reconstruction of truth for Case* 1 *for the eikonal equation. Top row: EKI reconstruction using random draws. Second row: TEKI reconstruction using random draws. Third row: EKI reconstruction using the KL basis. Second row: TEKI reconstruction using the KL basis.*

and the source term f defined as

$$f(x_1, x_2) = \begin{cases} 0 & \text{if } \ 0 \leq x_2 \leq \frac{4}{6}, \\ 137 & \text{if } \ \frac{4}{6} \leq x_2 \leq \frac{5}{6}, \\ 274 & \text{if } \ \frac{5}{6} \leq x_2 \leq 1. \end{cases}$$

The inverse problem is concerned with the recovery of $u = \log(\kappa)$ from mollified pointwise linear functionals of the form $G_j(u) = l_j(u)$, with l_j denoting mollified pointwise observation on a regular grid. The results that follow have no commentary because the phenomena exhibited are identical to what we see for the eikonal equation. For simplicity we keep to the same setting as subsection 4.3, where our results are presented in Figures 9, 10, 11, 12, 13, and 14.

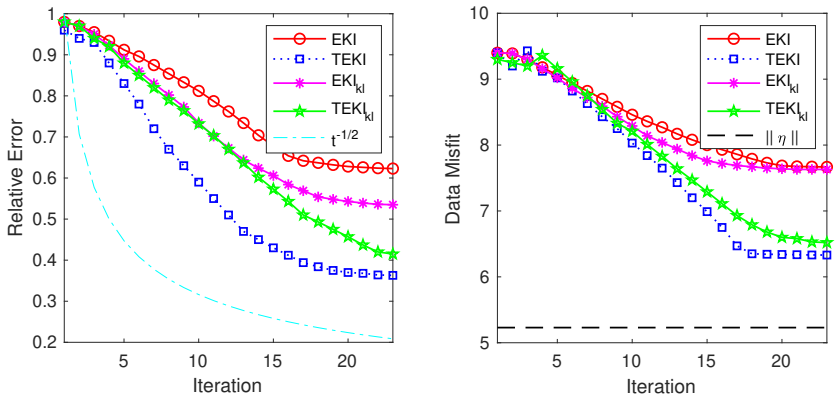

FIG. 4. *Case* 1. *Relative errors and data misfits for the eikonal equation.*

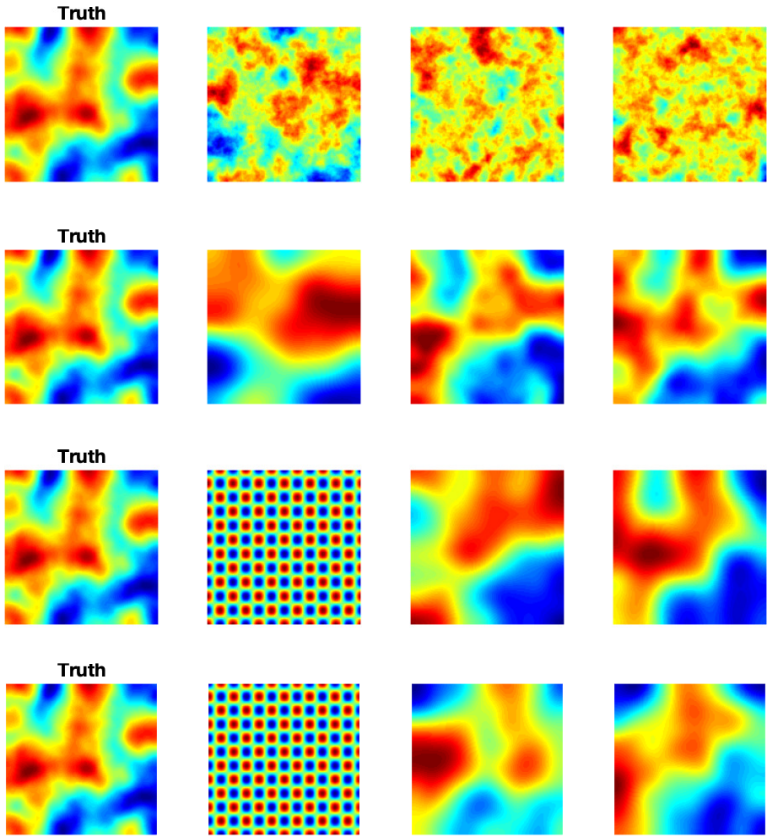

FIG. 5. *Reconstruction of truth for Case* 3 *for the eikonal equation. Top row: EKI reconstruction using random draws. Second row: TEKI reconstruction using random draws. Third row: EKI reconstruction using the KL basis. Second row: TEKI reconstruction using the KL basis.*

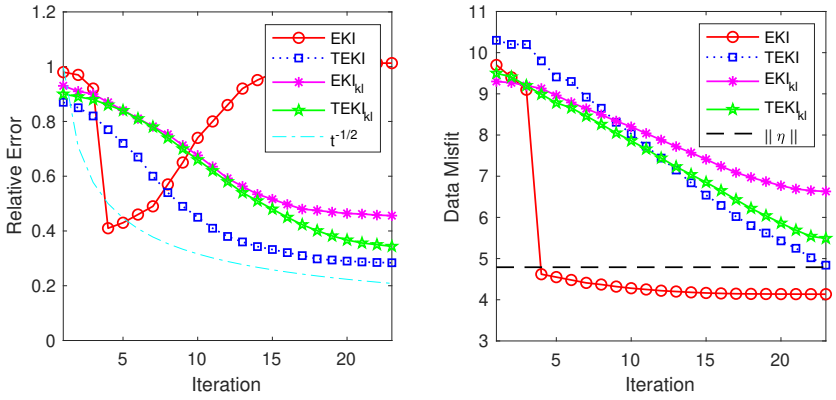

FIG. 6. *Case 2. Relative errors and data misfits for the eikonal equation.*

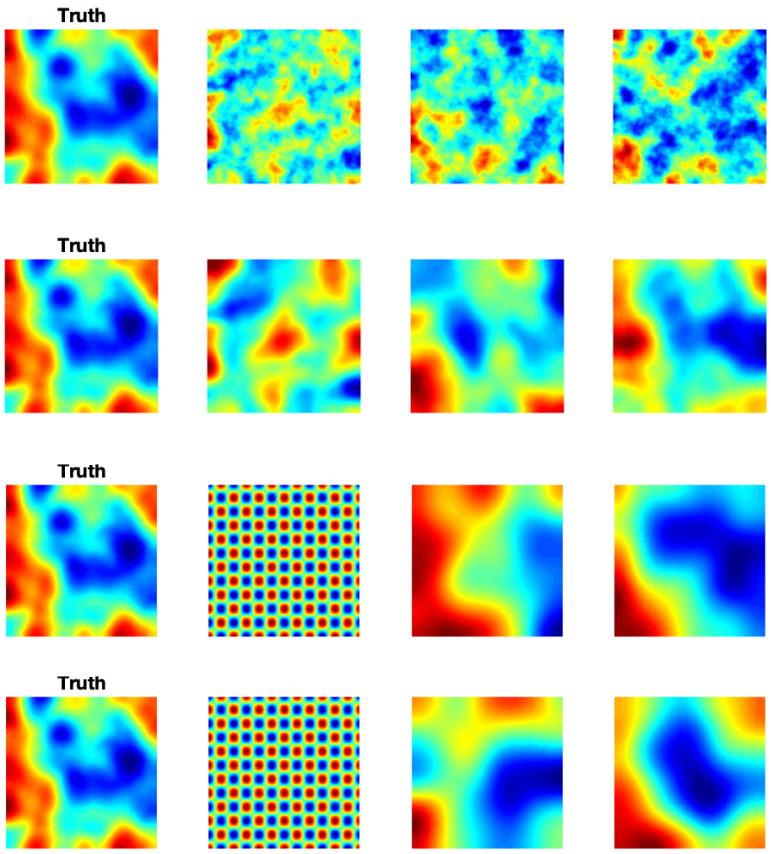

FIG. 7. *Reconstruction of truth for Case 3 for the eikonal equation. Top row: EKI reconstruction using random draws. Second row: TEKI reconstruction using random draws. Third row: EKI reconstruction using the KL basis. Second row: TEKI reconstruction using the KL basis.*

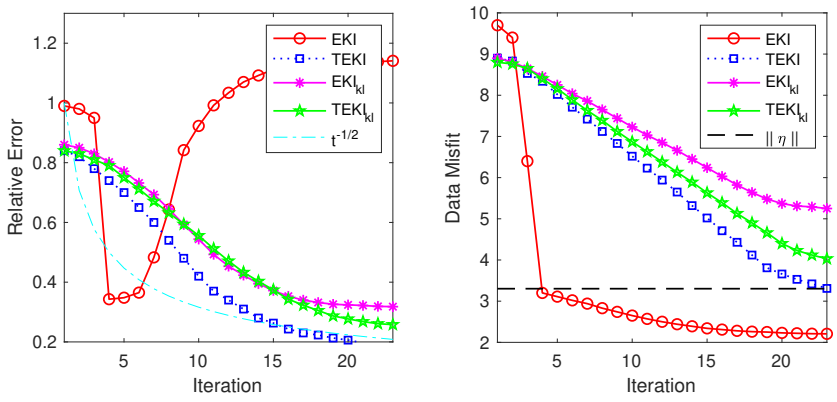

FIG. 8. *Case* 3. *Relative errors and data misfits for the eikonal equation.*

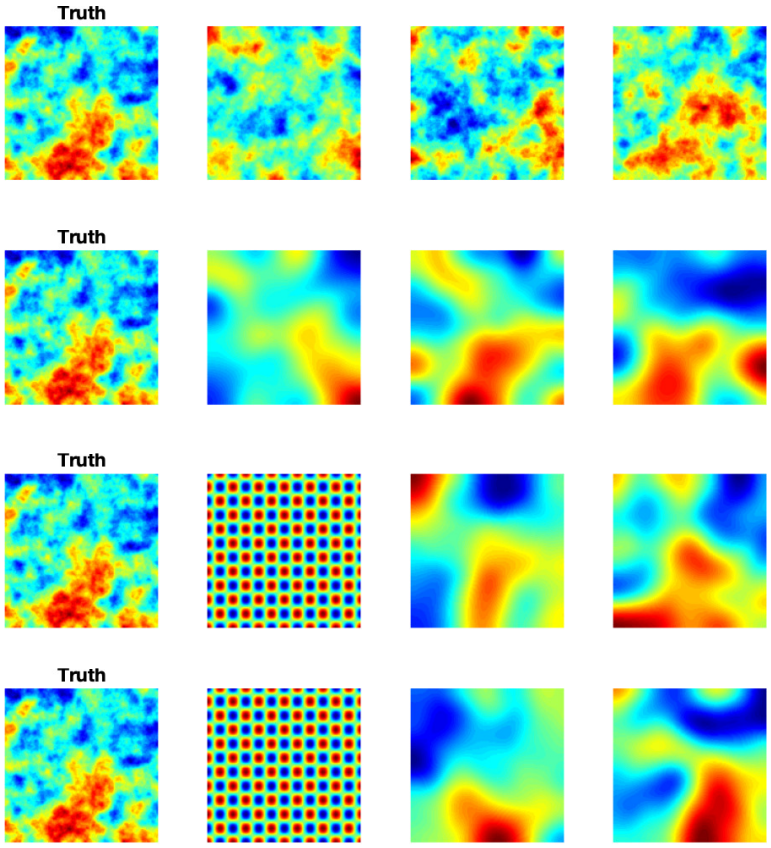

FIG. 9. *Reconstruction of truth for Case* 1 *for Darcy flow. Top row: EKI reconstruction using random draws. Second row: TEKI reconstruction using random draws. Third row: EKI reconstruction using the KL basis. Second row: TEKI reconstruction using the KL basis.*

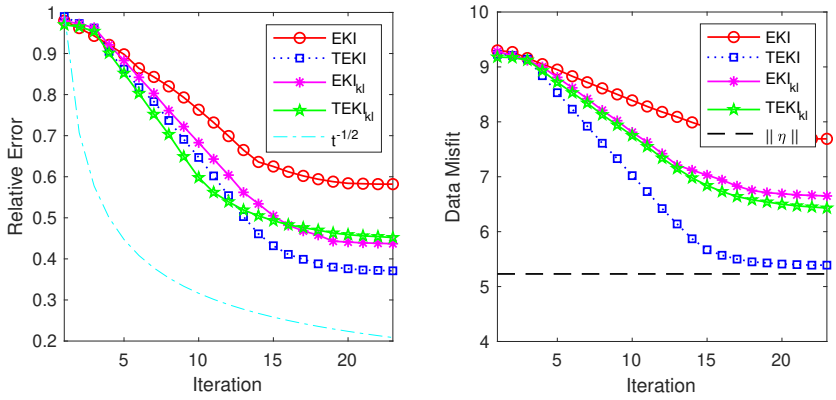

FIG. 10. *Case* 1. *Relative errors and data misfits for Darcy flow.*

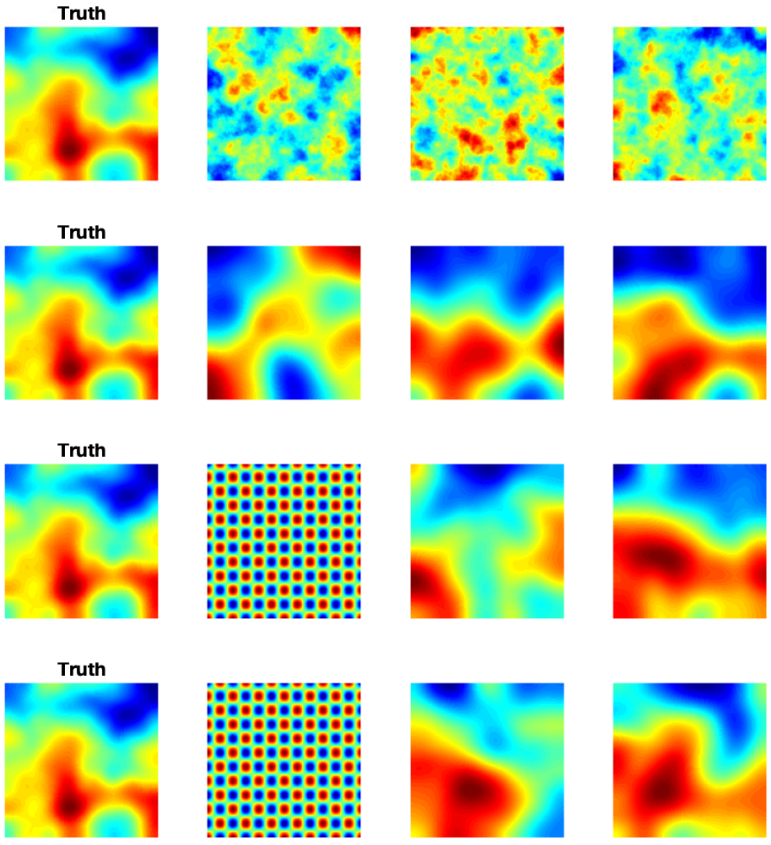

FIG. 11. *Reconstruction of truth for Case* 2 *for Darcy flow. Top row: EKI reconstruction using random draws. Second row: TEKI reconstruction using random draws. Third row: EKI reconstruction using the KL basis. Second row: TEKI reconstruction using the KL basis.*

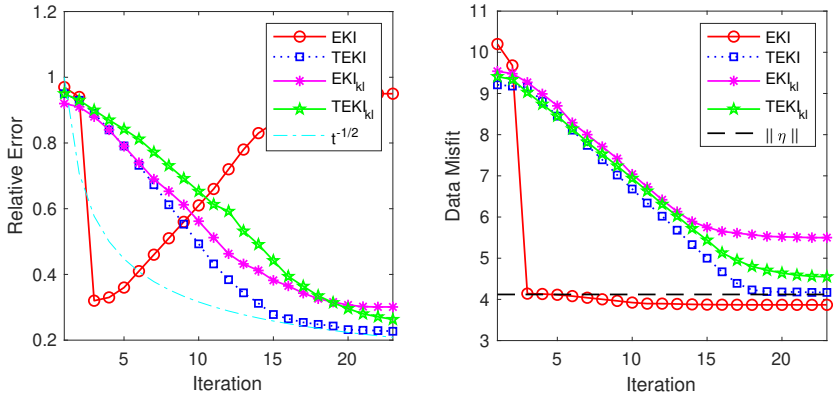

FIG. 12. *Case* 2. *Relative errors and data misfits for Darcy flow.*

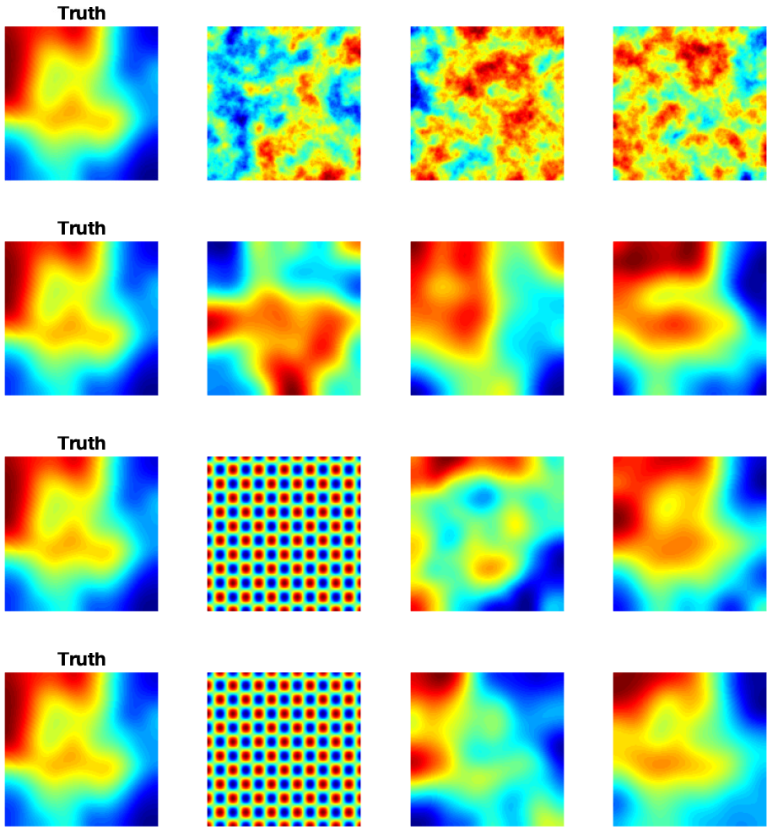

FIG. 13. *Reconstruction of truth for Case* 3 *for Darcy flow. Top row: EKI reconstruction using random draws. Second row: TEKI reconstruction using random draws. Third row: EKI reconstruction using the KL basis. Second row: TEKI reconstruction using the KL basis.*

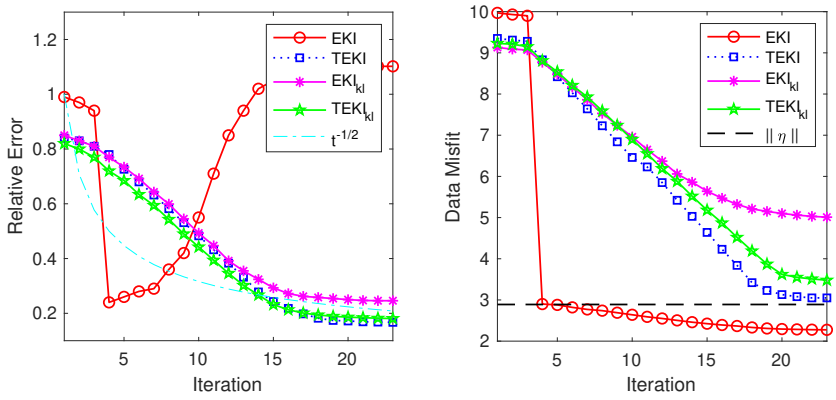

FIG. 14. *Case* 3. *Relative errors and data misfits for Darcy flow.*

5. Conclusions. Regularization is a central idea in optimization and statistical inference problems. In this work we considered adapting EKI methods to allow for Tikhonov regularization, leading to the TEKI methodology. Inclusion of this Tikhonov regularizer within EKI leads demonstrably to improved reconstructions of our unknown; we have shown this on an inverse eikonal equation and porous medium equation, using both random draws from the prior and the eigenfunctions of the KL basis to initialize the ensemble methods. We also derived a continuous time limit of TEKI and studied its properties, including showing the existence of the TEKI flow and its long-time behavior. In particular we showed that the TEKI flow always reaches consensus—ensemble members collapse on one another. Despite the theoretical and empirical improvements documented here, stemming from the use of TEKI over EKI, our concluding message is not that use of TEKI leads to the perfect algorithm for EKI. Rather we believe that the paper demonstrates that Tikhonov regularization can be incorporated into ensemble inversion methods in a simple fashion, and we provide evidence that doing so is worth considering; it can lead to improved reconstructions and smaller misfits, as well as having some theoretical advantages. Furthermore, we observe that the use of TEKI-related ideas is central to the ensemble Kalman sampler introduced in [17].

There are several potentially fruitful new directions one can consider which stem from this work; we outline a number of them:

- The inclusion of regularization in this paper was specific to the case of the Cameron–Martin space and hence Tikhonov-like Sobolev regularization. It would be of interest to generalize to the regularizers of other forms, such as L_1 and total variation penalties [3, 13].
- Understanding EKI as an optimizer is important, specifically in terms of how effective it is in comparison with other derivative-free optimization methods. Using the analysis tools developed in [7] could be helpful in this context. A related question to this direction is to assess the performance of this methodology for sampling, using the methodology introduced in [17].
- It would of interest to see how the techniques discussed in [5], where hierarchical EKI is introduced, could be improved by use of TEKI. The analysis presented here could be extended to the hierarchical setting.

- Related to hierarchical techniques discussed, one could treat the regularization parameter $\boldsymbol{\lambda}$ as a further unknown in our inverse problem. As this can be seen as a scaling factor in the covariance, it could be treated as an amplitude factor, in the usual way presented through Whittle–Matérn priors [38], and learned hierarchically as in [5]. Alternatively, it might be of interest to study the adaptation of other standard statistical techniques for estimation of $\boldsymbol{\lambda}$ to this inverse problem setting [3, 13, 44]. This is current work in progress [6].
- It is possible to impose convex constraints directly into EKI; see [2]. However, nonconvex constraints present difficulties in the framework described in that paper, as nonuniqueness may arise in the optimization problems to be solved at each step of the algorithm. Nonconvex equality constraints could be imposed by using the methods in this paper to impose them in a relaxed form. A constraint set defined by the equation $W(u) = 0$ could be approximately imposed by appending (2.4a)–(2.4b) with the equation $W(u) + \eta_3 = 0$ and choosing η_3 to be a Gaussian with small variance.
- Our long-time analysis has been developed in the finite-dimensional setting only; a natural extension would be to develop an infinite-dimensional theory, with assumptions guided by applications of importance for ensemble methods.

Acknowledgments. The authors are grateful to Vanessa Styles (University of Sussex) for providing a solver for the eikonal equation and Marco Iglesias (University of Nottingham) for providing a solver for the Darcy flow model, used, respectively, in [11, 20], and guidance on its use.

REFERENCES

[1] S. Agapiou, M. Burger, M. Dashti, and T. Helin, *Sparsity-promoting and edge-preserving maximum a posteriori estimators in non-parametric Bayesian inverse problems*, Inverse Problems, 34 (2018), 045002.

[2] D. J. Albers, P-A. Blancquart, M. E. Levine, E. E. Seylabi, and A. M. Stuart, *Ensemble Kalman methods with constraints*, Inverse Problems, 35 (2019), 095007.

[3] M. Benning and M. Burger, *Modern regularization methods for inverse problems*, Acta Numer., 27 (2018), pp. 1–111.

[4] D. Blömker, C. Schillings, and P. Wacker, *A strongly convergent numerical scheme from ensemble Kalman inversion*, SIAM J. Numer. Anal., 56 (2018), pp. 2537–2562, https://doi.org/10.1137/17M1132367.

[5] N. K. Chada, M. A. Iglesias, L. Roininen, and A. M. Stuart, *Parameterizations for ensemble Kalman inversion*, Inverse Problems, 32 (2018), 055009.

[6] N. K. Chada, C. Schillings, X. T. Tong, and S. Weissman, *Adaptive Learning in Stochastic Optimization: Applications to Ensemble Kalman Inversion*, manuscript.

[7] N. K. Chada and X. T. Tong, *Convergence Acceleration of Ensemble Kalman Inversion in Nonlinear Settings*, preprint, https://arxiv.org/abs/1911.02424, 2019.

[8] Y. Chen and D. S. Oliver, *Ensemble randomized maximum likelihood method as an iterative ensemble smoother*, Math. Geosci., 44 (2012), pp. 1–26.

[9] M. Dashti, K. J. H. Law, A. M. Stuart, and J. Voss, *MAP estimators and their consistency in Bayesian non-parametric inverse problems*, Inverse Problems, 29 (2013), 095017.

[10] M. Dashti and A. M. Stuart, *The Bayesian approach to inverse problems*, in Handbook of Uncertainty Quantification, Springer, Cham, 2016, pp. 1–118.

[11] C. M. Elliott, K. Deckelnick, and V. Styles, *Numerical analysis of an inverse problem for the eikonal equation*, Numer. Math., 119 (2011), pp. 245–269.

[12] A. A. Emerick and A. C. Reynolds, *Investigation of the sampling performance of ensemble-based methods with a simple reservoir model*, Comput. Geosci., 17 (2013), pp. 325–350.

[13] H. W. Engl, K. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Math. Appl. 375, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.

[14] O. G. Ernst, B. Sprungk, and H.-J. Starkloff, *Analysis of the ensemble and polynomial chaos Kalman filters in Bayesian inverse problems*, SIAM/ASA J. Uncertain. Quantif., 3

(2015), pp. 823–851, https://doi.org/10.1137/140981319.
[15] G. Evensen, *Data Assimilation: The Ensemble Kalman Filter*, Springer, Berlin, 2009.
[16] G. Evensen, *The ensemble Kalman filter: Theoretical formulation and practical implementation*, Ocean Dyn., 53 (2003), pp. 343–367.
[17] A. Garbuno-Inigo, F. Hoffmann, W. Li, and A. M. Stuart, *Interacting Langevin Diffusions: Gradient Structure and Ensemble Kalman Sampler*, preprint, https://arxiv.org/abs/1903.08866, 2019.
[18] T. Helin and M. Burger, *Maximum a posteriori probability estimates in infinite-dimensional Bayesian inverse problems*, Inverse Problems, 31 (2015), 085009.
[19] M. A. Iglesias, *A regularizing iterative ensemble Kalman method for PDE-constrained inverse problems*, Inverse Problems, 32 (2016), 025002.
[20] M. A. Iglesias, *Iterative regularization for ensemble data assimilation in reservoir models*, Comput. Geosci., 19 (2015), pp. 177–212.
[21] M. A. Iglesias, K. J. H. Law, and A. M. Stuart, *Ensemble Kalman methods for inverse problems*, Inverse Problems, 29 (2013), 045001.
[22] M. A. Iglesias, K. J. H. Law, and A. M. Stuart, *Evaluation of Gaussian approximations for data assimilation in reservoir models*, Comput. Geosci., 17 (2013), pp. 851–885.
[23] J. Kaipio and E. Somersalo, *Statistical and Computational Inverse problems*, Springer-Verlag, New York, 2004.
[24] D. Kelly, A. J. Majda, and X. T. Tong, *Concrete ensemble Kalman filters with rigorous catastrophic filter divergence*, Proc. Natl. Acad. Sci. USA, 112 (2016), pp. 10589–10594.
[25] D. T. Kelly, K. J. H. Law, and A. M. Stuart, *Well-posedness and accuracy of the ensemble Kalman filter in discrete and continuous time*, Nonlinearity, 27 (2014), pp. 2579–2604.
[26] N. Kovachi and A. M. Stuart, *Ensemble Kalman inversion: A derivative-free technique for machine learning tasks*, Inverse Problem, 35 (2019), 095005.
[27] K. J. H. Law and A. M. Stuart, *Evaluating data assimilation algorithms*, Mon. Weather Rev., 140 (2012), pp. 37–57.
[28] K. J. H. Law, A. M. Stuart, and K. Zygalakis, *Data Assimilation: A Mathematical Introduction*, Texts Appl. Math. 62, Springer, Cham, 2015.
[29] F. Le Gland, V. Monbet, and V. D. Tran, *Large sample asymptotics for the ensemble Kalman filter*, in The Oxford Handbook of Nonlinear Filtering, Oxford University Press, Oxford, UK, 2011, pp. 598–631.
[30] M. S. Lehtinen, L. Paivarinta, and E. Somersalo, *Linear inverse problems for generalised random variables*, Inverse Problems, 5 (1989), pp. 599–612.
[31] G. Li and A. C. Reynolds, *Iterative ensemble Kalman filters for data assimilation*, SPE J., 14 (2009), pp. 496–505.
[32] H-C. Lie and T. J. Sullivan *Equivalence of weak and strong modes of measures on topological vector spaces*, Inverse Problems, 34 (2018), 115013.
[33] P.-L. Lions, *Generalized Solutions of Hamilton-Jacobi Equations*, Pitman, Boston, 1982.
[34] D. M. Livings, S. L. Dance, and N. K. Nichols, *Unbiased ensemble square root filters*, Phys. D, 237 (2008), pp. 1021–1081.
[35] D. Oliver, A. C. Reynolds, and N. Liu, *Inverse Theory for Petroleum Reservoir Characterization and History Matching*, 1st ed., Cambridge University Press, Cambridge, UK, 2008.
[36] R. Pinnau, C. Totzeck, O. Tse, and S. Martin, *A consensus-based model for global optimization and its mean-field limit*, Math. Models Methods Appl. Sci., 27 (2017), pp. 183–204.
[37] S. Reich, *Data assimilation: The Schrödinger perspective*, Acta Numer., 28 (2019), pp. 635–711.
[38] L. Roininen, J. M. J. Huttunen, and S. Lasanen, *Whittle-Matérn priors for Bayesian statistical inversion with applications in electrical impedance tomography*, Inverse Probl. Imaging, 8 (2014), pp. 561–586.
[39] C. Schillings and A. M. Stuart, *Analysis of the ensemble Kalman filter for inverse problems*, SIAM J. Numer. Anal., 55 (2017), pp. 1264–1290, https://doi.org/10.1137/16M105959X.
[40] J. A. Sethian, *Level Set Methods and Fast Marching Methods*, Cambridge Monogr. Appl. Comput. Math. 3, Cambridge University Press, Cambridge, UK, 1999.
[41] A. M. Stuart, *Inverse problems: A Bayesian perspective*, Acta Numer., 19 (2010), pp. 451–559.
[42] X. T. Tong, *Performance analysis of local ensemble Kalman filter*, J. Nonlinear Sci., 28 (2018), pp. 1397–1442.

[43] X. T. Tong, A. J. Majda, and D. Kelly, *Nonlinear stability of the ensemble Kalman filter with adaptive covariance inflation*, Commun. Math. Sci., 14 (2016), pp. 1283–1313.
[44] G. Wahba, *A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem*, Ann. Statist., 13 (1985), pp. 1378–1402.
[45] M. Zupanski, M. I. Navon, and D. Zupanski, *The maximum likelihood ensemble filter as a non-differentiable minimization algorithm*, Q. J. R. Meteorol. Soc., 134 (2008), pp. 1039–1050.